

Bayesian Minimal Description Lengths for Multiple Changepoint Detection

Yingbo Li^{*†}, Robert Lund[†] and Hewa A. Priyadarshani[†]

Abstract

This paper develops a new class of flexible minimum description length (MDL) procedures for multiple changepoint detection. Existing MDL approaches, which are penalized likelihoods, use data description length information principles to construct penalties that depend on both the number of changepoints and the lengths of the series' segments. While MDL methods have yielded promising results in time series changepoint problems, state-of-the-art MDL approaches are not flexible enough to incorporate domain experts' knowledge that some times are more likely to be changepoints. Furthermore, current MDL methods do not readily handle multivariate series where changepoints can occur in some, but not necessarily all component series. The Bayesian MDL method developed in this paper provides a general framework to account for various prior knowledge, which substantially increases changepoint detection powers. Asymptotically, our estimated multiple changepoint configuration is shown to be consistent. Our method is motivated by a climate application, to identify mean shifts in monthly temperature records. In addition to autocorrelation and seasonal means, our method takes into account metadata, which is a record of station relocations and gauge changes, thus permitting study of documented and undocumented changepoint times in tandem. The multivariate extension allows maximum and minimum temperatures to be jointly examined.

^{*}Corresponding author, e-mail: ybli@clemson.edu.

[†]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634

Keywords: breakpoints, segmentation, structural breaks, empirical Bayes, time series, vector autoregression.

1 Introduction

Changepoints (i.e., structural breaks) are times in a time ordered record $\mathbf{X}_{1:N} = (X_1, \dots, X_N)'$ at which the data shift in some manner. The goal of a retrospective multiple changepoint analysis is to estimate the number of changepoints m and their locations $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)'$, $1 < \tau_1 < \dots < \tau_m \leq N$. Estimating the optimal changepoint configuration $(\hat{m}, \hat{\boldsymbol{\tau}})$ is a model selection problem, and can be conducted by optimizing a penalized likelihood.

In information theory, a description length (code length) is the number of binary storage units required to transmit a random number or code. The minimum description length (MDL) principle ([Rissanen 1989](#)) states that given the observed data, the model with the shortest code length is optimal. MDL methods have been successfully applied in segmentation problems ([Lee 2000](#); [Aue et al. 2014](#)). From a penalized likelihood perspective, MDL penalties are the sum of penalties (i.e., code lengths) of all unknown model parameters. Since MDL penalties are more than multiples of the number of model parameters, they are believed superior to AIC and BIC. Simulations support this conclusion (see [Lu et al. \(2010\)](#) and Section 5 here).

In the multiple changepoint literature, the seminal work by [Davis et al. \(2006\)](#) develops an MDL penalty for piecewise autoregressive (AR) processes. Following the automatic rules that (i) the code length (penalty) of an unbounded positive integer I is $\log_2(I)$, (ii) the code length of a positive integer bounded above by U is $\log_2(U)$, and (iii) the code length of the MLE of a real-valued parameter estimated by N observations is $\log_2(N)/2$, [Davis et al. \(2006\)](#) propose

$$\log(m) + (m + 1) \log(N) + \sum_{r=1}^{m+1} \log p_r + \sum_{r=1}^{m+1} \frac{p_r + 2}{2} \log N_r \quad (1)$$

as the MDL penalty after replacing all \log_2 logarithms by natural logarithms (this does not alter the minimum). Here, (1) is the sum of penalties on all model parameters, with m and

the AR orders p_r in segments $r = 1, \dots, m+1$ being unbounded integers, the segment lengths $N_r = \tau_r - \tau_{r-1}$ being integers bounded by N , and the AR parameters (the AR coefficients, the mean, and the variance) of the r th segment being real-valued of dimension $p_r + 2$.

MDL methods are considered state-of-the-art for multiple changepoint problems (Chan et al. 2014), and are extended to various time series processes, including GARCH (Davis et al. 2008), periodic AR (Lu et al. 2010), and threshold AR (Yau et al. 2015) (see also Shao and Zhang (2010) and Preuss et al. (2015) for CUSUM and spectral-based approaches to the multiple changepoint problem). For rapid computation, pre-screening approaches such as group Lasso (Chan et al. 2014) and likelihood ratio scan statistics (Yau and Zhao 2015) can be used to minimize the MDL score. However, issues exist with these automatic MDL methods: the term $\log(m)$ is infinite when $m = 0$, the methods do not readily handle cases where prior information is available, and multivariate cases where only a subset of the components change at a changepoint time are not accommodated. This paper remedies these issues.

Our motivation lies in climate homogenization problems (Caussinus and Mestre 2004; Menne and Williams Jr 2005, 2009; Lu et al. 2010; Li and Lund 2012). With temperature time series, abrupt changes are often attributed to artificial causes such as station relocations, gauge changes, or observer changes. Without changepoint adjustments, temperature trend analyses can be misleading (see Lu and Lund (2007) and the references therein). Metadata station history logs, which document the times of physical changes in the station, are often available. Although metadata records are notoriously incomplete, and not all documented times induce actual shifts in time series, climatologists believe that incorporating metadata increases changepoint detection power. One goal of this paper is to account for the prior knowledge in metadata — these times are much more likely to induce mean shifts. Simulation results (e.g., Figure 4) demonstrate that incorporating the metadata significantly increases the detection powers at the documented times that are true changepoints, and meanwhile decreases the average false positive detection rate.

This paper also considers multivariate series, specifically monthly averages of daily max-

imum (Tmax) and minimum temperatures (Tmin). [Davis et al. \(2006\)](#) assume that each changepoint affects all component series in the multivariate setting simultaneously. This is not the case for Tmax and Tmin series: a station move to a drier location can simultaneously increase daytime highs and reduce nighttime lows, while a station move to a more sheltered location may decrease daytime highs but leave nighttime lows unaltered. While changepoints occurring in both Tmax and Tmin at the same time (concurrent shifts) are more likely, changepoints in either series by themselves can also occur. For current MDLs, it is not clear whether a concurrent shift should be counted as one or two changes in the penalty; moreover, definition of the segment lengths in the penalty becomes nebulous.

By revisiting basic MDL principles ([Hansen and Yu 2001](#)), the connection between code lengths of model parameters and their prior distributions from the Bayesian perspective becomes clear. Bayesian model selection procedures, where the model that optimizes the posterior probability is usually selected, can be loosely viewed as penalized likelihoods: in the posterior distributions, the prior densities act as penalties. Multiple changepoint approaches have also been devised using Bayesian model selection ideas ([Barry and Hartigan 1993](#); [Chib 1998](#); [Girón et al. 2007](#); [Giordani and Kohn 2008](#); [Fearnhead and Vasileiou 2009](#); [Hannart and Naveau 2012](#); [Du et al. 2015](#)).

This paper aims to develop a new class of flexible MDL methods. Inspired by the Bayesian model selection literature ([Clyde and George 2004](#)), we reformulate the multiple changepoint configuration $\boldsymbol{\tau}$ as a vector of zero/one indicators, where the t th component indicates whether time t is a changepoint or not. This enables natural construction of flexible prior distributions on the changepoint configuration, with straightforward hyper-parameter elicitation. Our MDL method is termed a Bayesian MDL (BMDL), because it is closely related to empirical Bayes model selection approaches, and thus permits the use of stochastic search algorithms such as Markov chain Monte Carlo (MCMC) to achieve efficient computation. While our main focus is to improve and generalize the conventional MDL changepoint detection approaches, to the best of our knowledge, this paper is the first among Bayesian model selection works to show

asymptotic model selection consistency with autocorrelated observations.

In the rest of this paper, Section 2 reviews MDL principles, which are needed to modify the aforementioned automatic code length rules. Section 3 develops a BMDL for a univariate time series that accounts for a metadata record. For temperature homogenization problems, our method is tailored to detect mean shifts, allowing for a seasonal cycle and autocorrelated errors. Under infill asymptotics, we show that the BMDL is consistent, and also compare it with the existing MDL. Section 4 extends our method to the bivariate case to study Tmax and Tmin series jointly. Section 5 presents a simulation study to demonstrate the efficacy of our method. Section 6 presents an application to 114 years of monthly temperatures from Tuscaloosa, Alabama. Comments close the paper in Section 7. Technical details and theorem proofs are contained in the supplementary material.

2 A Brief Review of MDL

In data transmission, to reduce storage costs, one wants to assign shorter (longer) code lengths to common (rare) outcomes. Competing probability models can be compared by their code lengths; the true data generating distribution (i.e., the true model) should have the shortest expected code length.

For a discrete random variable X taking values in \mathcal{X} with probability mass function $f(\cdot)$, [Shannon \(1948\)](#) states that the encoding with code length $\mathcal{L}(X) = -\log_2 f(X)$ has the shortest expected code length. For example, if X is uniformly distributed over $\{1, 2, \dots, U\}$, then its MDL is $\mathcal{L}(X) = -\log_2(1/U) = \log_2(U)$. This leads to the automatic code length rule (ii) for bounded integers, which is mentioned in [Section 1](#). For the automatic rule (i), when the positive integer $X \in \{1, 2, \dots\}$ is unbounded, the code length $\log_2(X)$ is implied under an improper power law distribution $f(X) \propto 1/X$.

If \mathbf{X} is a k -dimensional continuous variable with density function $f(\cdot)$, after discretizing each dimension into equal cells of size δ (often viewed as the machine precision), the discrete

case can be mimicked to obtain $\mathcal{L}(\mathbf{X}) = -\log_2[f(\mathbf{X})\delta^k] = -\log_2 f(\mathbf{X}) - k\log_2(\delta)$. Because k and δ do not vary with \mathbf{X} , the term $-k\log_2(\delta)$ does not affect comparison between different \mathbf{X} and is often omitted. One can substitute the natural logarithm for the base two logarithm — this does not affect model comparisons since $\log_2(x)/\log(x)$ is constant in x .

Now suppose that a dataset $\mathbf{X} = (X_1, \dots, X_n)'$, believed to be generated from a certain parametric model \mathcal{M} with density $f(\mathbf{X} \mid \theta, \mathcal{M})$, is to be transmitted along with a possibly unknown parameter $\theta \in \Theta$. As reviewed in [Hansen and Yu \(2001\)](#), two types of MDL approaches, the two-part MDL and the mixture MDL, are commonly used.

2.1 Two-part MDLs

The two-part MDL, also called the two-stage MDL, considers the transmission of \mathbf{X} and θ in two steps. If both the sender and receiver know θ , the MDL of \mathbf{X} is $\mathcal{L}(\mathbf{X} \mid \theta, \mathcal{M}) = -\log f(\mathbf{X} \mid \theta, \mathcal{M})$. Here, notations such as $\mathcal{L}(\cdot \mid \cdot)$ are analogous to the usual conditional distribution notations to emphasize dependence. Should θ also be unknown to the receiver, an additional cost of $\mathcal{L}(\theta \mid \mathcal{M})$ is incurred in transmitting it. Hence, the two-part MDL becomes $\mathcal{L}(\mathbf{X}, \theta \mid \mathcal{M}) = \mathcal{L}(\mathbf{X} \mid \theta, \mathcal{M}) + \mathcal{L}(\theta \mid \mathcal{M})$.

Suppose $\mathcal{L}(\mathbf{X}, \theta \mid \mathcal{M})$ is minimized at $\hat{\theta}$, an estimator of θ based on the data \mathbf{X} . If θ is a k -dimensional continuous parameter and $\hat{\theta}$ is a \sqrt{n} -consistent estimator of θ , then one can set the machine precision to be $\delta = c/\sqrt{n}$, where c is a positive constant. Under a uniform encoder $\pi(\theta \mid \mathcal{M}) \propto 1$, the code length needed to transmit θ (including $\hat{\theta}$) is hence $\mathcal{L}(\theta \mid \mathcal{M}) = -\log \pi(\theta \mid \mathcal{M}) - k\log(c/\sqrt{n}) = k\log(n)/2 - k\log(c)$, which does not depend on θ . Hence, the maximum likelihood estimator (MLE) minimizes $\mathcal{L}(\mathbf{X}, \theta \mid \mathcal{M})$, and the two-part MDL coincides with the BIC ([Schwarz 1978](#)). In fact, $\hat{\theta}$ need not be the MLE; any \sqrt{n} -consistent estimator is justifiable. After dropping the constant term $k\log(c)$, the code length $\mathcal{L}(\hat{\theta} \mid \mathcal{M}) = k\log(n)/2$ agrees with the automatic code length rule (iii) that is previously mentioned in [Section 1](#).

If there exists a discrete set of candidate models, to account for model uncertainty, the

two-part MDL can be modified to include an additional code length for the model \mathcal{M} , i.e.,

$$\mathcal{L}(\mathbf{X}, \hat{\theta}, \mathcal{M}) = \mathcal{L}(\mathbf{X} \mid \hat{\theta}, \mathcal{M}) + \mathcal{L}(\hat{\theta} \mid \mathcal{M}) + \mathcal{L}(\mathcal{M}), \quad (2)$$

where $\hat{\theta}$ is model dependent, $\mathcal{L}(\mathcal{M}) = -\log \pi(\mathcal{M})$, and $\pi(\mathcal{M})$ is the prior distribution over the model space. The model with the smallest MDL in (2) is deemed optimal.

All existing automatic MDL methods for multiple changepoint detection are based on two-part MDLs. However, for a finite sample size n , the two-part MDL is problematic when the dimension of θ changes across models, as in the multiple changepoint case. Consider a setting of two competing models \mathcal{M}_1 and \mathcal{M}_2 , whose parameters θ_j are k_j -dimensional continuous parameters, for $j = 1, 2$, and $k_1 \neq k_2$. Model \mathcal{M}_1 is favored if $\mathcal{L}(\mathbf{X}, \hat{\theta}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\theta}_2, \mathcal{M}_2)$ is negative; otherwise, model \mathcal{M}_2 is favored. The code length difference for the parameters $\mathcal{L}(\hat{\theta}_1 \mid \mathcal{M}_1) - \mathcal{L}(\hat{\theta}_2 \mid \mathcal{M}_2)$ contains the term $(k_1 - k_2)[\log(n) - 2\log(c)]/2$. This term, and hence also $\mathcal{L}(\mathbf{X}, \hat{\theta}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\theta}_2, \mathcal{M}_2)$, could be either positive or negative depending on n and c . One cannot judge either model superior without knowledge of c . Of course, this issue does not conflict with the asymptotic consistency of BIC or automatic MDLs: as n increases, $\log(n)$ dominates the constant $\log(c)$. Mixture MDLs considered next do not suffer from such a problem under finite n .

2.2 Mixture MDLs

For the model \mathcal{M} , suppose that θ has a prior distribution with density $\pi(\theta \mid \mathcal{M})$. The marginal likelihood of \mathbf{X} averages the likelihood $f(\mathbf{X} \mid \theta, \mathcal{M})$ under this prior distribution:

$$f(\mathbf{X} \mid \mathcal{M}) = \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \mathcal{M}) d\theta.$$

According to [Hansen and Yu \(2001\)](#), the mixture MDL is defined based on the marginal likelihood: $\mathcal{L}(\mathbf{X} \mid \mathcal{M}) = -\log f(\mathbf{X} \mid \mathcal{M}) = -\log \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \mathcal{M}) d\theta$. If the prior of θ

depends on an unknown hyper-parameter ψ , then a two-part MDL can be used to account for the additional cost needed to transmit ψ . In this case, the overall mixture MDL, for any \sqrt{n} -consistent estimator of ψ , is $\mathcal{L}(\mathbf{X}, \hat{\psi} \mid \mathcal{M}) = -\log \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \hat{\psi}, \mathcal{M}) d\theta + \mathcal{L}(\hat{\psi} \mid \mathcal{M})$.

The mixture MDL for the model \mathcal{M} is thus $\mathcal{L}(\mathbf{X}, \hat{\psi}, \mathcal{M}) = \mathcal{L}(\mathbf{X}, \hat{\psi} \mid \mathcal{M}) + \mathcal{L}(\mathcal{M})$, which is related to empirical Bayes (EB) approaches. If the prior probabilities of two models are the same, i.e., $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$, and the hyper-parameter ψ is transmitted under the uniform encoder $\pi(\psi \mid \mathcal{M}_j) \propto 1$ for $j = 1, 2$, then the logarithm of their Bayes factor (Kass and Raftery 1995) $\text{BF}_{\mathcal{M}_2:\mathcal{M}_1}$ equals the difference of their mixture MDLs, $\mathcal{L}(\mathbf{X}, \hat{\psi}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\psi}_2, \mathcal{M}_2)$. Similarly, in EB settings, while the estimator $\hat{\psi}$ is often chosen to maximize the marginal likelihood $f(\mathbf{X} \mid \psi, \mathcal{M})$, other estimators can be used (Carlin and Louis 2000).

3 Bayesian MDL

Consider a univariate time series $\mathbf{X}_{1:N}$ with a seasonal mean cycle of a fundamental period T . For monthly data, $T = 12$. A model with autoregressive errors describing this situation is

$$X_t = s_{v(t)} + \mu_{r(t)} + \epsilon_t, \quad \epsilon_t = \sum_{j=1}^p \phi_j \epsilon_{t-j} + Z_t. \quad (3)$$

Here, $v(t) = t - T \lfloor (t-1)/T \rfloor$ is the season corresponding to time t , where $\lfloor x \rfloor$ is the largest integer less than or equal to x . The seasonal means $\mathbf{s} = (s_1, \dots, s_T)'$ are unknown. The errors $\{\epsilon_t\}_{t=1}^N$ are a causal zero mean AR process of a known order p . The AR coefficients $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$ and the white noise variance $\text{Var}(Z_t) = \sigma^2$ are assumed unknown. In climate applications, monthly averaged temperatures are approximately normally distributed (Wilks 2011). Hence, in further likelihood computations, Gaussianity is assumed.

Suppose m changepoints located at the times $\tau_1 < \dots < \tau_m$ partition the observation times $\{1, \dots, N\}$ into $m + 1$ distinct regimes (segments), where the series mean minus the seasonal mean, $\mu_{r(t)}$, changes across regimes. To avoid trite work with edge effects of the

autoregression, we assume that no changepoints occur during the first p observations. For notation, let $\tau_0 = 1$ and $\tau_{m+1} = N + 1$. The regime indicator $r(t)$ in (3) satisfies $r(t) = r$ when $\tau_{r-1} \leq t < \tau_r$. To ensure parameter identifiability, μ_1 is taken as zero; hence, $E[X_t] = s_{v(t)}$ when t lies in the first regime. The other regime means $\boldsymbol{\mu} = (\mu_2, \dots, \mu_{m+1})'$ are unknown.

Following Li and Lund (2015), we write the multiple changepoint configuration $(m; \boldsymbol{\tau})$ as an $(N - p)$ -dimensional vector of zero/one indicators: $\boldsymbol{\eta} = (\eta_{p+1}, \dots, \eta_N)'$. Here, $\eta_t = 1$ means that time t is a changepoint; $\eta_t = 0$ means that time t is not a changepoint. The total number of changepoints in model $\boldsymbol{\eta}$ is thus $m = \sum_{t=p+1}^N \eta_t$.

Our main idea is to apply the mixture MDL to the continuous parameter $\boldsymbol{\mu}$, whose dimension varies across models, and use the two-part MDL for the parameters $\mathbf{s}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}$, and the model $\boldsymbol{\eta}$. In the rest of this section, Section 3.1 introduces our prior choices on $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, Section 3.2 derives the BMDL formula (14), Section 3.3 discusses computational strategies, and finally Section 3.4 studies asymptotic properties.

3.1 Prior specifications

Our prior distribution for the changepoint model $\boldsymbol{\eta}$ assumes that, in the absence of metadata, each time t has an equal probability ρ of being a changepoint, independently of all other times, i.e., $\eta_t \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\rho)$, for $t = p + 1, \dots, N$. This independent Bernoulli prior has been used in previous Bayesian multiple changepoint detection works (Chernoff and Zacks 1964; Yao 1984; Barry and Hartigan 1993). Then $\tau_r \mid \tau_{r-1} \sim \text{Geometric}(\rho)$ for $r = 1, \dots, m$, which is used in Fearnhead and Vasileiou (2009) and is a special case of the Negative Binomial prior in Hannart and Naveau (2012). The uniform prior $\pi(\boldsymbol{\eta}) \propto 1$ in Du et al. (2015) is a special case of the Bernoulli prior with $\rho = 0.5$. For applications where knowledge beyond the metadata is unavailable, an iid prior on $\{\eta_t\}$ seems reasonable. In other applications, $\pi(\boldsymbol{\eta})$ can have different success probabilities in different regimes (Chib 1998); correlation across different changepoint times can also be achieved using Ising priors (Li and Zhang 2010).

To account for uncertainty in the success probability ρ , a hyper-prior is placed on it.

Barry and Hartigan (1993) allow ρ to have a uniform prior on the interval $(0, \rho_0)$, where $\rho_0 < 1$. For additional flexibility, we use the Beta distribution $\rho \sim \text{Beta}(a, b)$, where $a, b > 0$ are fixed hyper-parameters. Beta-Binomial hierarchical priors are common in Bayesian model selection (Scott and Berger 2010). Due to Beta-Binomial conjugacy, the marginal prior density of $\boldsymbol{\eta}$ has the closed form

$$\pi(\boldsymbol{\eta}) = \int_0^1 \left[\prod_{t=p+1}^N \pi(\eta_t \mid \rho) \right] \pi(\rho) d\rho = \frac{\beta(a+m, b+N-p-m)}{\beta(a, b)}, \quad (4)$$

where $\beta(\cdot, \cdot)$ denotes the Beta function.

For hyper-parameter choices, an objective Bayesian option (Girón et al. 2007) is $a = b = 1$. In this case, $\pi(\boldsymbol{\eta}) = \left[\binom{N-p}{m} (N-p+1) \right]^{-1}$, which implies that marginally, m has a uniform prior on the set $\{0, 1, \dots, N-p\}$, and conditional on m , all models containing m changepoints have the same prior probabilities. The Beta-Binomial prior can be tuned to accommodate subjective knowledge from domain experts (Giordani and Kohn 2008). For temperature homogenization, Mitchell (1953) estimates an average of six station relocations and gauge changes per century in United States temperature series; this long-term rate is 0.005 changepoints per month and can be produced with $a = 1$ and $b = 199$; with these parameters, $E(\rho) = a/(a+b) = 0.005$.

This prior is now modified to accommodate metadata. Suppose that during the times $\{p+1, \dots, N\}$, there are $N^{(2)}$ documented times (times listed in the metadata) and $N^{(1)} = N - p - N^{(2)}$ undocumented times. For notation, all quantities superscripted with (1) refer to undocumented times; quantities superscripted with (2) refer to documented times. Following Li and Lund (2015), we posit that the undocumented times have a Beta-Binomial($a, b^{(1)}$) prior, and independently, the documented times have a Beta-Binomial($a, b^{(2)}$) prior. To make the metadata times more likely to induce true mean shifts, we impose $b^{(1)} > b^{(2)}$ such that

$$E[\rho^{(1)}] = \frac{a}{a+b^{(1)}} < \frac{a}{a+b^{(2)}} = E[\rho^{(2)}].$$

For monthly data, default values are $a = 1$, $b^{(1)} = 239$, and $b^{(2)} = 47$, making $E(\rho^{(1)}) = 0.0042$ and $E(\rho^{(2)}) = 0.0208$; *a priori*, a documented time is roughly five times as likely to be a changepoint as an undocumented time. One may change the values of $b^{(1)}$ and $b^{(2)}$ to reflect different prior beliefs. The sensitivity analysis in [Li and Lund \(2015\)](#) suggests that changepoint detection results are relatively stable under a range of $E(\rho^{(2)})/E(\rho^{(1)})$ values.

Following (4) and writing Beta integrals via their Gamma function representations, a changepoint configuration $\boldsymbol{\eta}$ with $m^{(2)}$ documented changepoints and $m^{(1)}$ undocumented changepoints ($m = m^{(1)} + m^{(2)}$) has a marginal prior density (up to a normalizing constant)

$$\pi(\boldsymbol{\eta}) \propto \prod_{k=1}^2 \Gamma(a + m^{(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{(k)}).$$

For a changepoint model with $m > 0$ changepoints, priors for the m -dimensional regime means $\boldsymbol{\mu}$ are posited to have independent normal prior distributions:

$$\boldsymbol{\mu} \mid \sigma^2, \boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \nu \sigma^2 \mathbf{I}_m).$$

Here, ν is a pre-specified non-negative parameter that is relatively large (making the variances of the regime means large multiples of the white noise variances). Similar to the sensitivity analysis in [Du et al. \(2015\)](#), our experience suggests that model selection results are stable under a wide range of ν values. Our default takes $\nu = 5$.

In fact, $\pi(\boldsymbol{\mu})$ can be any zero mean continuous distribution. For example, if mean shifts are expected to be large, heavy-tailed distributions such as the Student- t may be preferable. When $\boldsymbol{\mu}$ cannot be tractably integrated out, inferences can be based on Laplace approximations or posterior sampling with a reversible-jump MCMC ([Green 1995](#)). Due to conjugacy under Gaussian likelihoods, the normal prior leads to closed form marginal likelihoods. Hence, for computational ease in the rest of this paper, normal regime mean priors are used.

3.2 The BMDL expressions

To derive the BMDL expression (14), we first develop the data likelihood, then integrate $\boldsymbol{\mu}$ out to obtain the mixture MDL, and finally add the two-part MDLs of the rest of the parameters.

Given a changepoint model $\boldsymbol{\eta}$, the sampling distribution (3) has the regression representation

$$\mathbf{X}_{1:N} = \mathbf{A}_{1:N}\mathbf{s} + \mathbf{D}_{1:N}\boldsymbol{\mu} + \boldsymbol{\epsilon}_{1:N}, \quad (5)$$

with $\mathbf{A}_{1:N} \in \mathbb{R}^{N \times T}$ and $\mathbf{D}_{1:N} \in \mathbb{R}^{N \times m}$ as seasonal and regime indicator matrices:

$$\begin{aligned} [\mathbf{A}_{1:N}]_{t,v} &= \mathbf{1}(\text{time } t \text{ is in season } v), \quad v = 1, \dots, T, \\ [\mathbf{D}_{1:N}]_{t,r-1} &= \mathbf{1}(\text{time } t \text{ is in regime } r), \quad r = 2, \dots, m+1, \end{aligned}$$

where $\mathbf{1}(A)$ denotes the indicator of the event A . In (5), the subscript $1:N$, or in general $t_1:t_2$, signifies that only rows t_1 through t_2 are used in the quantities. The normal white noises $\{Z_t\}$ in the AR process imply the distributional result $\boldsymbol{\epsilon}_{(p+1):N} - \sum_{j=1}^p \phi_j \boldsymbol{\epsilon}_{(p+1-j):(N-j)} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p})$, where \mathbf{I}_k denotes the $k \times k$ identity matrix. Now define

$$\tilde{\mathbf{X}} = \mathbf{X}_{(p+1):N} - \sum_{j=1}^p \phi_j \mathbf{X}_{(p+1-j):(N-j)}, \quad (6)$$

$$\tilde{\mathbf{A}} = \mathbf{A}_{(p+1):N} - \sum_{j=1}^p \phi_j \mathbf{A}_{(p+1-j):(N-j)}, \quad \tilde{\mathbf{D}} = \mathbf{D}_{(p+1):N} - \sum_{j=1}^p \phi_j \mathbf{D}_{(p+1-j):(N-j)}, \quad (7)$$

and observe that

$$\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p}). \quad (8)$$

Note that all terms superscripted with \sim depend on the unknown AR parameter $\boldsymbol{\phi}$. To avoid AR edge effects, a likelihood conditional on the initial observations $\mathbf{X}_{1:p}$ is used. In the change of variable computations, the Jacobian $|\partial(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu})/\partial \mathbf{X}_{(p+1):N}| = 1$ and the likelihood

has the multivariate normal form

$$f(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = (2\pi\sigma^2)^{-\frac{N-p}{2}} e^{-\frac{1}{2\sigma^2}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu})'(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu})}.$$

Innovation forms of the likelihood (Brockwell and Davis 1991) could be used if one wants a moving-average or long-memory component in $\{\epsilon_t\}$.

We now obtain a BMDL for the changepoint model $\boldsymbol{\eta}$. If $m > 0$, we first use the mixture MDL on $\boldsymbol{\mu}$. The marginal likelihood, after integrating $\boldsymbol{\mu}$ out, has the closed form

$$\begin{aligned} f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) &= \int_{\mathbb{R}^m} f(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) \pi(\boldsymbol{\mu} \mid \sigma^2, \boldsymbol{\eta}) d\boldsymbol{\mu} \\ &= (2\pi\sigma^2)^{-\frac{N-p}{2}} \nu^{-\frac{m}{2}} \left| \tilde{\mathbf{D}}' \tilde{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s})' \tilde{\mathbf{B}}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s})}, \end{aligned}$$

where the notation has

$$\tilde{\mathbf{B}} = \mathbf{I}_{N-p} - \tilde{\mathbf{D}} \left(\tilde{\mathbf{D}}' \tilde{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right)^{-1} \tilde{\mathbf{D}}'. \quad (9)$$

If the parameters \mathbf{s} , σ^2 , and $\boldsymbol{\phi}$ are known, the mixture MDL is simply $\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = -\log f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta})$.

The two-part MDL is used to quantify the cost of transmitting the parameters \mathbf{s} , σ^2 , and $\boldsymbol{\phi}$. The optimal \mathbf{s} and σ^2 that minimize the mixture MDL have the closed forms:

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s}} \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = (\tilde{\mathbf{A}}' \tilde{\mathbf{B}} \tilde{\mathbf{A}})^{-1} (\tilde{\mathbf{A}}' \tilde{\mathbf{B}} \tilde{\mathbf{X}}), \quad (10)$$

$$\hat{\sigma}^2 = \arg \min_{\sigma^2} \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = \frac{1}{N-p} \tilde{\mathbf{X}}' \left[\tilde{\mathbf{B}} - \tilde{\mathbf{B}} \tilde{\mathbf{A}} (\tilde{\mathbf{A}}' \tilde{\mathbf{B}} \tilde{\mathbf{A}})^{-1} \tilde{\mathbf{A}}' \tilde{\mathbf{B}} \right] \tilde{\mathbf{X}}. \quad (11)$$

These estimators depend on $\boldsymbol{\phi}$; however, the $\boldsymbol{\phi}$ that minimizes $\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{\eta})$ is intractable. In general, likelihood estimators for autoregressive models do not have closed forms. Hence, simple Yule-Walker moment estimators, which are asymptotically most efficient and \sqrt{n} -consistent under the true changepoint model, are used. There is little difference between moment and likelihood estimators for autoregressions (Brockwell and Davis 1991).

In the linear model (5), the ordinary least squares residuals are

$$\boldsymbol{\epsilon}_{1:N}^{\text{ols}} = (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_{1:N} \ \mathbf{D}_{1:N}]})\mathbf{X}_{1:N}, \quad (12)$$

where $\mathcal{P}_{[\mathbf{A}_{1:N} \ \mathbf{D}_{1:N}]}$ is the orthogonal projection matrix onto the linear space spanned by the columns of $\mathbf{A}_{1:N}$ and $\mathbf{D}_{1:N}$. The sample autocovariance of the residuals at lag $h = 0, 1, \dots, p$ are $\hat{\gamma}(h) = N^{-1} \sum_{t=h+1}^N \epsilon_t^{\text{ols}} \epsilon_{t-h}^{\text{ols}}$. The Yule-Walker estimator of $\boldsymbol{\phi}$ is $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p$, where $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \dots, \hat{\gamma}(p))'$ and $\hat{\boldsymbol{\Gamma}}_p$ is a $p \times p$ matrix whose (i, j) th entry is $\hat{\gamma}(|i - j|)$. This matrix is invertible whenever the data are non-constant (Brockwell and Davis 1991). Next, the Yule-Walker estimator $\hat{\boldsymbol{\phi}}$ is substituted for $\boldsymbol{\phi}$ in $\tilde{\mathbf{X}}$, $\tilde{\mathbf{A}}$, $\tilde{\mathbf{D}}$, and $\tilde{\mathbf{B}}$. The resulting quantities are denoted by $\hat{\mathbf{X}}$, $\hat{\mathbf{A}}$, $\hat{\mathbf{D}}$, and $\hat{\mathbf{B}}$, respectively. In particular, $\hat{\mathbf{X}}$ contains estimated one-step-ahead prediction residuals (innovations).

By (2), the Bayesian MDL for transmitting the data $\mathbf{X}_{(p+1):N}$, the model $\boldsymbol{\eta}$, and its parameters is

$$\begin{aligned} \text{BMDL}(\boldsymbol{\eta}) &= \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}) + \mathcal{L}(\hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) + \mathcal{L}(\boldsymbol{\eta}) \\ &= -\log f(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}) - \log \pi(\boldsymbol{\eta}). \end{aligned} \quad (13)$$

The second equality holds because under a uniform encoder $\pi(\mathbf{s}, \sigma^2, \boldsymbol{\phi}) \propto 1$, the two-part MDL $\mathcal{L}(\hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = (T + 1 + p) \log(N - p)/2$ is constant across models and hence can be omitted. Therefore, for a model with $m > 0$ changepoints, its BMDL is (up to a constant)

$$\begin{aligned} \text{BMDL}(\boldsymbol{\eta}) &= \frac{N - p}{2} \log(\hat{\sigma}^2) + \frac{m}{2} \log(\nu) + \frac{1}{2} \log \left(\left| \hat{\mathbf{D}}' \hat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right| \right) \\ &\quad - \sum_{k=1}^2 \log [\Gamma(a + m^{(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{(k)})]. \end{aligned} \quad (14)$$

For a model with no changepoints ($m = 0$), denoted by $\boldsymbol{\eta}_\emptyset$, the above procedure needs modification. Since $\boldsymbol{\eta}_\emptyset$ does not involve $\boldsymbol{\mu}$, the mixture MDL step can be skipped. As \mathbf{D} has

no columns, $\tilde{\mathbf{B}}$ in (9) is reduced to \mathbf{I}_{N-p} , and hence (11) still holds. With the convention that the determinant of a 0×0 matrix is unity, $\log \left(\left| \hat{\mathbf{D}}' \hat{\mathbf{D}} + \mathbf{I}_m / \nu \right| \right) = 0$. Therefore, (14) also holds for $\boldsymbol{\mu}$. This resolves the issue of evaluating $\log(m)$ at $m = 0$ with other MDL methods.

3.3 BMDL optimization

The optimal changepoint model $\hat{\boldsymbol{\eta}}$ is selected as the one with the smallest BMDL score. However, exhaustively searching the changepoint configuration space is formidable, since the total number of models, 2^{N-p} , is extremely large. To overcome this hurdle, stochastic search algorithms such as the genetic algorithm are used (Davis et al. 2006; Lu et al. 2010), which efficiently explore the model space to only visit a relatively small number of promising models.

The connection between BMDL and empirical Bayes (EB) allows us to borrow MCMC model search algorithms that are commonly used in Bayesian model selection. Note that the BMDL under model $\boldsymbol{\eta}$ represented in (13) is equivalent to the negative logarithm of an EB estimator of the posterior probability of $\boldsymbol{\eta}$:

$$p_{\text{EB}}(\boldsymbol{\eta} \mid \mathbf{X}_{(p+1):N}) \propto \pi(\boldsymbol{\eta}) \int_{\mathbb{R}^m} f\left(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}\right) \pi(\boldsymbol{\mu} \mid \hat{\sigma}^2, \boldsymbol{\eta}) d\boldsymbol{\mu}.$$

As our BMDL formula (14) is tractable, Bayesian stochastic model search algorithms can be used; see García-Donato and Martínez-Beneito (2013) and the references therein.

Here, we modify the Metropolis-Hastings algorithm in George and McCulloch (1997) by intertwining two types of proposals: a component-wise flipping at a random location and a simple random swapping between a changepoint and a non-changepoint. This algorithm is described in detail in Li and Lund (2015) and can be implemented by an R package provided by the authors.

3.4 Asymptotic properties of the BMDL

Infill asymptotics assume that the number of observations between all changepoints tends to infinity and have been widely adopted to study consistency of multiple changepoint detection procedures (Davis et al. 2006; Davis and Yau 2013; Du et al. 2015). Under infill asymptotics, a relative changepoint configuration with m changepoints is denoted by $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)'$, where $0 < \lambda_1 < \dots < \lambda_m < 1$. Here, time is scaled to $[0, 1]$ by mapping time t to t/N . For the edges, set $\lambda_0 = 0$ and $\lambda_{m+1} = 1$. For a given N , the r th changepoint location τ_r can be recovered from $\boldsymbol{\lambda}$ via $\tau_r = \lfloor \lambda_r N \rfloor$. The length of the r th regime, $N_r = \lfloor \lambda_r N \rfloor - \lfloor \lambda_{r-1} N \rfloor$, satisfies $\lim_{N \rightarrow \infty} N_r/N = \lambda_r - \lambda_{r-1}$, for $r = 1, \dots, m+1$. For any $\boldsymbol{\lambda}$, no changepoints occur in $\{1, \dots, p\}$ when N is large.

Suppose that the true relative changepoint configuration is $\boldsymbol{\lambda}^0 = (\lambda_1^0, \dots, \lambda_{m^0}^0)'$, where true parameter values are superscripted with zero. Our goal is to identify $\boldsymbol{\lambda}^0$ over many candidate models. In fact, for a (fixed) large integer M , all relative changepoint configurations in

$$\boldsymbol{\Lambda} = \{\boldsymbol{\lambda} : 0 \leq m \leq M, \min_{r=1,2,\dots,m+1} \lambda_r - \lambda_{r-1} \geq d\}$$

are considered, where d is a small positive constant, smaller than $\lambda_r^0 - \lambda_{r-1}^0$ for all $r = 1, \dots, m^0 + 1$. We assume that $m^0 \leq M$; hence, $\boldsymbol{\lambda}^0 \in \boldsymbol{\Lambda}$. Between the true model $\boldsymbol{\lambda}^0$ and any other model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, the pairwise difference of their BMDLs in (14) is used to decide which model is favorable.

Theorem 1. *For any relative changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, if $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^0$, then as $N \rightarrow \infty$,*

$$BMDL(\boldsymbol{\lambda}) - BMDL(\boldsymbol{\lambda}^0) \xrightarrow{P} \infty.$$

More specifically, if all relative changepoints in $\boldsymbol{\lambda}^0$ are contained in $\boldsymbol{\lambda}$, then the above BMDL difference is $O_P(\log N)$; otherwise, it is $O_P(N)$.

The proof of Theorem 1 is provided in the supplementary material. This theorem shows

that asymptotically, the true relative changepoint model $\boldsymbol{\lambda}^0$ achieves the smallest BMDL in probability among all competing models in $\boldsymbol{\Lambda}$. As an implication of the result, it is possible to consistently identify the true changepoint configuration in the limit.

We now compare our BMDL with an MDL derived by the automatic code length rules (i) - (iii), for the same multiple mean shifts problem in (3):

$$\text{MDL}(\boldsymbol{\eta}) = \frac{N-p}{2} \log(\hat{\sigma}_{\nu=\infty}^2) + \frac{1}{2} \sum_{j=2}^{m+1} \log(N_r) + \log(m+1) + (m+1) \log(N-p). \quad (15)$$

The first term in (15) is the negative logarithm of the maximum likelihood, where the estimator of σ^2 follows (11) with $\nu = \infty$ and hence is denoted by $\hat{\sigma}_{\nu=\infty}^2$. The rest of the terms in (15) are the two-part MDLs of the regime means μ_2, \dots, μ_{m+1} , the number of changepoints m (an ad-hoc fix to the $\log(m=0)$ problem is to use $m+1$ in the logarithm), and the regime lengths N_1, \dots, N_{m+1} , respectively. The two-part MDLs of the global parameters \mathbf{s} , σ^2 , and ϕ are constants and hence omitted. An MDL for the AR order p is not needed since p is assumed known.

Under a given relative changepoint model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, both the BMDL (14) and the automatic MDL (15) increase linearly with N , i.e., $\text{BMDL}(\boldsymbol{\lambda}) = O_P(N)$ and $\text{MDL}(\boldsymbol{\lambda}) = O_P(N)$. The following proposition states that when metadata is not available, the model selection performance of (14) and (15) are asymptotically the same.

Proposition 1. *For any two relative changepoint models $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda}$, as $N \rightarrow \infty$,*

$$\text{BMDL}(\boldsymbol{\lambda}_1) - \text{BMDL}(\boldsymbol{\lambda}_2) = \text{MDL}(\boldsymbol{\lambda}_1) - \text{MDL}(\boldsymbol{\lambda}_2) + O_P(1).$$

An implication of Proposition 1 is that the model selection consistency results in Theorem 1 also hold for the automatic MDL, which confirms the asymptotic results in Davis et al. (2006) and Davis and Yau (2013).

Proposition 1 is shown by comparing the large sample performance of the corresponding

terms in (14) and (15). In the BMDL expression (14), all but the last term arise from the mixture MDL. The first term $(N - p) \log(\hat{\sigma}^2)/2$ measures the goodness-of-fit. By Lemma 2 in the supplementary material, the difference between the first terms in (14) and (15) is bounded,

$$\frac{N - p}{2} \log(\hat{\sigma}^2) - \frac{N - p}{2} \log(\hat{\sigma}_{\nu=\infty}^2) = O_P(1).$$

The second term in (14) is $O_P(1)$. By Lemma 4 in the supplementary material, the third term in (14) is $O_P(\log N)$, and

$$\frac{1}{2} \log \left(\left| \hat{\mathbf{D}}' \hat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right| \right) = \frac{1}{2} \sum_{j=2}^{m+1} \log(N_r) + O_P(1),$$

which interestingly suggests that the mixture MDL in (14) contains a built-in penalty on $\boldsymbol{\mu}$ that performs similarly to the two-part MDL penalty on $\boldsymbol{\mu}$ in (15). The last term in (14) is the penalty on the changepoint configuration $\boldsymbol{\lambda}$. Assuming no metadata, we denote it by $\mathcal{L}_B(\boldsymbol{\lambda}) = -\log[\Gamma(a + m) \Gamma(b + N - p - m)]$. Its automatic MDL counterpart is the sum of the last two terms in (15), denoted by $\mathcal{L}_A(\boldsymbol{\lambda}) = \log(m + 1) + (m + 1) \log(N - p)$. Note that both $\mathcal{L}_B(\boldsymbol{\lambda})$ and $\mathcal{L}_A(\boldsymbol{\lambda})$ only depend on m , not the relative changepoint locations. For two competing models $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda}$ with m_1, m_2 changepoints, respectively, when N is large, $\mathcal{L}_A(\boldsymbol{\lambda}_2) - \mathcal{L}_A(\boldsymbol{\lambda}_1) = (m_2 - m_1) \log(N) + O_P(1)$. Stirling's formula implies the same result for the BMDL counterpart: $\mathcal{L}_B(\boldsymbol{\lambda}_2) - \mathcal{L}_B(\boldsymbol{\lambda}_1) = (m_2 - m_1) \log(N) + O_P(1)$.

Therefore, without metadata, the BMDL (14) and the automatic MDL (15) perform similarly under large samples. Section 5 confirms this result via simulation examples, and meanwhile demonstrates that when metadata is available and incorporated, the BMDL significantly increases changepoint detection accuracy.

4 Extensions to Bivariate Time Series

Following the same procedure, this section develops the BMDL expression for bivariate time series, allowing changepoints to occur in either or both component series. This procedure can be easily generalized to examine multivariate series of more than two components.

To model Tmax and Tmin series jointly, we concatenate the observed series via $\mathbf{X}_{1:N} = (\mathbf{X}'_{1:N,1}, \mathbf{X}'_{1:N,2})' \in \mathbb{R}^{2N}$, where $\mathbf{X}_{1:N,i} = (X_{1,i}, \dots, X_{N,i})'$ is the record for Tmax ($i = 1$) or Tmin ($i = 2$). Each time in $\{p+1, \dots, N\}$ is allowed to be a changepoint in either the Tmax or Tmin series, or both. A multiple changepoint configuration is denoted by $\boldsymbol{\eta} = (\boldsymbol{\eta}'_1, \boldsymbol{\eta}'_2)'$, where $\boldsymbol{\eta}_i = (\eta_{p+1,i}, \dots, \eta_{N,i})' \in \{0, 1\}^{N-p}$ is defined as in the univariate case. Given a bivariate changepoint model $\boldsymbol{\eta}$, series i has $m_i = \sum_{t=p+1}^N \eta_{t,i}$ changepoints. As in the univariate case, the seasonal means are denoted by $\mathbf{s}_i = (s_{1,i}, \dots, s_{T,i})' \in \mathbb{R}^T$; regime means are denoted by $\boldsymbol{\mu}_i = (\mu_{2,i}, \dots, \mu_{m_i+1,i})' \in \mathbb{R}^{m_i}$. The seasonal and regime indicator matrices $\mathbf{A}_{1:N,i} \in \mathbb{R}^{N \times T}$ and $\mathbf{D}_{1:N,i} \in \mathbb{R}^{N \times m_i}$ are constructed analogously to their univariate counterparts.

The regression representation (5) holds for the bivariate case, with $\mathbf{s} = (\mathbf{s}'_1, \mathbf{s}'_2)'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2)'$, $\boldsymbol{\epsilon}_{1:N} = (\boldsymbol{\epsilon}'_{1:N,1}, \boldsymbol{\epsilon}'_{1:N,2})'$ denoting the concatenated seasonal means, regime means, and regression errors, respectively. The seasonal and regime indicator matrices have the block diagonal forms $\mathbf{A}_{1:N} = \text{diag}(\mathbf{A}_{1:N,1}, \mathbf{A}_{1:N,2})$ and $\mathbf{D}_{1:N} = \text{diag}(\mathbf{D}_{1:N,1}, \mathbf{D}_{1:N,2})$. Note that the seasonal indicators for Tmax and Tmin coincide, i.e., $\mathbf{A}_{1:N,1} = \mathbf{A}_{1:N,2}$, while $\mathbf{D}_{1:N,1}$ and $\mathbf{D}_{1:N,2}$ differ unless all changepoints are concurrent.

As Tmax and Tmin temperature data tend to fluctuate about the mean in tandem (positive correlation), the errors $\{\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \epsilon_{t,2})'\}$ need to be correlated across components. For this, a vector autoregressive model (VAR) of order p is employed:

$$\boldsymbol{\epsilon}_t = \sum_{j=1}^p \boldsymbol{\Phi}_j \boldsymbol{\epsilon}_{t-j} + \mathbf{Z}_t, \quad \text{Cov}(\mathbf{Z}_t) = \boldsymbol{\Sigma},$$

Here, $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p$ are 2×2 VAR coefficient matrices. The VAR model allows for correlation in time and between components.

As (8) holds after replacing $\sigma^2 \mathbf{I}_{N-p}$ with $\boldsymbol{\Sigma} \otimes \mathbf{I}_{N-p}$, the likelihood of $\mathbf{X}_{(p+1):N}$, conditional on the initial observations $\mathbf{X}_{1:p}$, is (up to a multiplicative constant)

$$f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}_{1:p}, \boldsymbol{\eta}) \propto |\boldsymbol{\Sigma}|^{-\frac{N-p}{2}} e^{-\frac{1}{2}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu})'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s} - \tilde{\mathbf{D}}\boldsymbol{\mu})}.$$

Here, \otimes denotes a Kronecker product and the terms $\tilde{\mathbf{X}}, \tilde{\mathbf{A}}, \tilde{\mathbf{D}}$ are modified by replacing ϕ_j with $\boldsymbol{\Phi}_j \otimes \mathbf{I}_{N-p}$ in (6) and (7), for $j = 1, \dots, p$.

4.1 Prior specifications

For $t = p + 1, \dots, N$, the indicator $\boldsymbol{\eta}_t = (\eta_{t,1}, \eta_{t,2})'$ takes values in one of the four categories: $(1, 1)'$, mean shifts in both Tmax and Tmin; $(1, 0)'$, a mean shift in Tmax but not in Tmin; $(0, 1)'$, a mean shift in Tmin but not in Tmax; and $(0, 0)'$, no mean shifts. As a natural extension of the Beta-Binomial prior, a Dirichlet-Multinomial prior is put on $\boldsymbol{\eta}_t$:

$$\boldsymbol{\eta}_t \mid \boldsymbol{\rho} \stackrel{\text{iid}}{\sim} \text{Multinomial}(1; \boldsymbol{\rho}), \quad \boldsymbol{\rho} \sim \text{Dirichlet}(\boldsymbol{\alpha}),$$

where $\boldsymbol{\rho} = (\rho_1, \dots, \rho_4)'$ are the probabilities of the four categories satisfying $\sum_{\ell=1}^4 \rho_\ell = 1$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_4)'$ are the Dirichlet parameters with $\alpha_\ell > 0$ for each $\ell = 1, \dots, 4$. Suppose that the changepoint configuration $\boldsymbol{\eta}$ has m_ℓ times in category ℓ . Due to Dirichlet-multinomial conjugacy, the marginal prior of $\boldsymbol{\eta}$ has a closed form after integrating out $\boldsymbol{\rho}^{(1)}$ and $\boldsymbol{\rho}^{(2)}$:

$$\pi(\boldsymbol{\eta}) \propto \prod_{k=1}^2 \prod_{\ell=1}^4 \Gamma(\alpha_\ell^{(k)} + m_\ell^{(k)}).$$

The choice of the hyper-parameter $\boldsymbol{\alpha}$ reflects our belief that concurrent changepoints are more likely to occur than an independent scenario. The ratios between the prior expectations satisfy $E(\rho_1) : E(\rho_2) : E(\rho_3) : E(\rho_4) = \alpha_1 : \alpha_2 : \alpha_3 : \alpha_4$. If changepoints in the Tmax and Tmin series at time t are independent events, then $\rho_1 = P(\eta_{t,1} = 1, \eta_{t,2} = 1) = P(\eta_{t,1} = 1)P(\eta_{t,2} = 1) = (\rho_1 + \rho_2)(\rho_1 + \rho_3)$. To encourage concurrent shifts, it is assumed that they

occur more often than in independent settings. This is done by choosing $\boldsymbol{\alpha}$ such that

$$E(\rho_1) = \frac{\alpha_1}{\sum_{\ell=1}^4 \alpha_\ell} > \frac{\alpha_1 + \alpha_2}{\sum_{\ell=1}^4 \alpha_\ell} \frac{\alpha_1 + \alpha_3}{\sum_{\ell=1}^4 \alpha_\ell} = E(\rho_1 + \rho_2)E(\rho_1 + \rho_3).$$

In addition, we match the probability of no changepoints with its counterpart in the univariate case, in terms of the prior mean, i.e., $\alpha_4 / \sum_{\ell=1}^4 \alpha_\ell = b/(a+b)$. After consulting climatologists, default hyper-parameters are set to $\boldsymbol{\alpha}^{(1)} = (3/7, 2/7, 2/7, 239)'$ and $\boldsymbol{\alpha}^{(2)} = (3/7, 2/7, 2/7, 47)'$ for monthly data.

To obtain the mixture MDL in a closed form, for a bivariate model with $m = m_1 + m_2 > 0$ changepoints, the regime means $\boldsymbol{\mu}$ again have independent normal priors

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \nu \operatorname{diag} \left(\underbrace{\sigma_1^2, \dots, \sigma_1^2}_{m_1}, \underbrace{\sigma_2^2, \dots, \sigma_2^2}_{m_2} \right),$$

where σ_1^2 and σ_2^2 are the diagonal entries of the white noise covariance $\boldsymbol{\Sigma}$.

4.2 The bivariate BMDL

For a bivariate model $\boldsymbol{\eta}$ with $m > 0$, the marginal likelihood, after integrating $\boldsymbol{\mu}$ out, has a closed form:

$$f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}_{1:p}, \boldsymbol{\eta}) \propto |\boldsymbol{\Sigma}|^{-\frac{N-p}{2}} |\boldsymbol{\Omega}|^{-\frac{1}{2}} \left| \tilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})\tilde{\mathbf{D}} + \boldsymbol{\Omega}^{-1} \right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s})' \tilde{\mathbf{B}}(\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\mathbf{s})},$$

where $\tilde{\mathbf{B}}$ is modified to

$$\tilde{\mathbf{B}} = (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p}) \times \left\{ \mathbf{I}_{2(N-p)} - \tilde{\mathbf{D}} \left[\tilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})\tilde{\mathbf{D}} + \boldsymbol{\Omega}^{-1} \right]^{-1} \tilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p}) \right\}.$$

The maximum marginal likelihood estimator $\tilde{\mathbf{s}}$ is unaltered from (10). However, after plugging $\hat{\mathbf{s}}$ back into the likelihood, the maximum likelihood estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p$ again do not have closed forms. Therefore, Yule-Walker estimators are used.

To find Yule-Walker estimators for the time series regression (5), generalized least squares residuals of the mean fit, denoted by $\boldsymbol{\epsilon}_{1:N}^{\text{gls}} = ((\boldsymbol{\epsilon}_{1:N,1}^{\text{gls}})', (\boldsymbol{\epsilon}_{1:N,2}^{\text{gls}})')' \in \mathbb{R}^{2N}$, are computed via

$$\boldsymbol{\epsilon}_{1:N}^{\text{gls}} = \left\{ \mathbf{I}_{2N} - \mathbf{G} \left[\mathbf{G}' \left(\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0)^{-1} \otimes \mathbf{I}_N \right) \mathbf{G} \right]^{-1} \mathbf{G}' \left(\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0)^{-1} \otimes \mathbf{I}_N \right) \right\} \mathbf{X}_{1:N},$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{1:N,1} & \mathbf{D}_{1:N,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{1:N,2} & \mathbf{D}_{1:N,2} \end{bmatrix}.$$

Here, $\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0) = N^{-1} \sum_{t=1}^N \boldsymbol{\epsilon}_t^{\text{ols}} (\boldsymbol{\epsilon}_t^{\text{ols}})'$ is a 2×2 covariance matrix of the ordinary (unweighted) least squares residuals $\boldsymbol{\epsilon}_t^{\text{ols}} = (\epsilon_{t,1}^{\text{ols}}, \epsilon_{t,2}^{\text{ols}})'$, where $\epsilon_{t,1}^{\text{ols}}$ and $\epsilon_{t,2}^{\text{ols}}$ are computed analogously to (12) with the design matrices $[\mathbf{A}_{1:N,1} \ \mathbf{D}_{1:N,1}]$ and $[\mathbf{A}_{1:N,2} \ \mathbf{D}_{1:N,2}]$, respectively. The sample autocovariances at lag $h = 0, 1, \dots, p$ of the generalized least squares residuals $\boldsymbol{\epsilon}_t^{\text{gls}} = (\epsilon_{t,1}^{\text{gls}}, \epsilon_{t,2}^{\text{gls}})'$, $t = 1, \dots, N$ are computed as $\hat{\boldsymbol{\Gamma}}(h) = N^{-1} \sum_{t=h+1}^N \boldsymbol{\epsilon}_t^{\text{gls}} (\boldsymbol{\epsilon}_{t-h}^{\text{gls}})'$. The Yule-Walker estimators in the bivariate setting are

$$\begin{pmatrix} \hat{\boldsymbol{\Phi}}_1, \dots, \hat{\boldsymbol{\Phi}}_p \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\Gamma}}(1), \dots, \hat{\boldsymbol{\Gamma}}(p) \end{pmatrix} \begin{bmatrix} \hat{\boldsymbol{\Gamma}}(0) & \hat{\boldsymbol{\Gamma}}(1) & \dots & \hat{\boldsymbol{\Gamma}}(p-1) \\ \hat{\boldsymbol{\Gamma}}(1)' & \hat{\boldsymbol{\Gamma}}(0) & \dots & \hat{\boldsymbol{\Gamma}}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\boldsymbol{\Gamma}}(p-1)' & \hat{\boldsymbol{\Gamma}}(p-2)' & \dots & \hat{\boldsymbol{\Gamma}}(0) \end{bmatrix}^{-1}$$

and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}}(0) - \sum_{j=1}^p \hat{\boldsymbol{\Phi}}_j \hat{\boldsymbol{\Gamma}}(j)'$.

After plugging $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Phi}}_1, \dots, \hat{\boldsymbol{\Phi}}_p$ back into the marginal likelihood, the terms $\tilde{\mathbf{X}}, \tilde{\mathbf{A}}, \tilde{\mathbf{D}}, \tilde{\mathbf{B}}, \boldsymbol{\Omega}$, which depend on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p$, are denoted by $\hat{\mathbf{X}}, \hat{\mathbf{A}}, \hat{\mathbf{D}}, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}$, respectively. Hence, the Bayesian MDL for $\boldsymbol{\eta}$ is (up to a constant)

$$\begin{aligned} \text{BMDL}(\boldsymbol{\eta}) = & \frac{N-p}{2} \log \left(|\hat{\boldsymbol{\Sigma}}| \right) + \frac{1}{2} \sum_{i=1}^2 m_i \log(\nu \hat{\sigma}_i^2) + \frac{1}{2} \log \left(\left| \hat{\mathbf{D}}' (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_{N-p}) \hat{\mathbf{D}} + \hat{\boldsymbol{\Omega}}^{-1} \right| \right) \\ & + \frac{1}{2} \hat{\mathbf{X}}' \left[\hat{\mathbf{B}} - \hat{\mathbf{B}} \hat{\mathbf{A}} \left(\hat{\mathbf{A}}' \hat{\mathbf{B}} \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}' \hat{\mathbf{B}} \right] \hat{\mathbf{X}} - \sum_{k=1}^2 \sum_{\ell=1}^4 \log \left[\Gamma \left(\alpha_{\ell}^{(k)} + m_{\ell}^{(k)} \right) \right]. \end{aligned}$$

Under the null model η_\emptyset , because $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}^{-1} \otimes \mathbf{I}_{N-p}$, with the convention that the determinant of a 0×0 matrix is unity, the above BMDL still holds.

5 Simulation Studies

This section studies changepoint detection performance under finite samples via simulation. Our simulation parameters are selected to roughly resemble the bivariate Tuscaloosa data, which will be studied in Section 6. Specifically, the bivariate error series $\{\epsilon_t\}$ follows a zero mean Gaussian VAR model with $p = 3$. The VAR parameters are taken as

$$\Phi_1 = \begin{pmatrix} 0.2 & 0.02 \\ 0.02 & 0.2 \end{pmatrix}, \Phi_2 = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}, \Phi_3 = \begin{pmatrix} 0.05 & 0.005 \\ 0.005 & 0.05 \end{pmatrix}, \Sigma = \begin{pmatrix} 9 & 2 \\ 2 & 9 \end{pmatrix}.$$

In each of 1000 independent runs, 50 year monthly Tmax and Tmin series ($N = 600$) are simulated with $m = 3$ changepoints in each series. For the Tmax series, mean shifts are placed at times 150, 300, and 450. The regime means have form $\mu_1 = (0, \Delta, 2\Delta, 3\Delta)'$ where $\Delta > 0$ will be varied. For the Tmin series, mean shifts are placed at times 150, 300, and 375. The regime means are $\mu_2 = (0, -\Delta, \Delta, 0)'$. Here, Tmax has monotonic “up, up, up” shifts of equal shift magnitudes; Tmin shifts in a “down, up, down” fashion and the second shift is twice as large as the other two shifts. The shifts at times 150 and 300 are concurrent in both series; the shift at time 150 increases Tmax and decreases Tmin.

Seasonal means are set to $\mathbf{s} = (0, 3, 10, 18, 26, 33, 36, 36, 31, 20, 8, 2)'$ in both series. Seasonal mean parameters are not critical, but the Δ parameter controlling the mean shift size is. Our detection powers will be reported under different signal to noise ratios, which we take as $\kappa = \Delta/\sigma$. We will examine $\kappa \in \{1, 1.5, 2\}$, where $\sigma = 3$. For metadata, a record containing four documented changes at the times 75, 150, 250, and 550 is posited. Among the documented times, only time 150 is a true changepoint.

A simulated series with $\kappa = 1.5$ is shown in Figure 1. Figure 2 shows the same series after

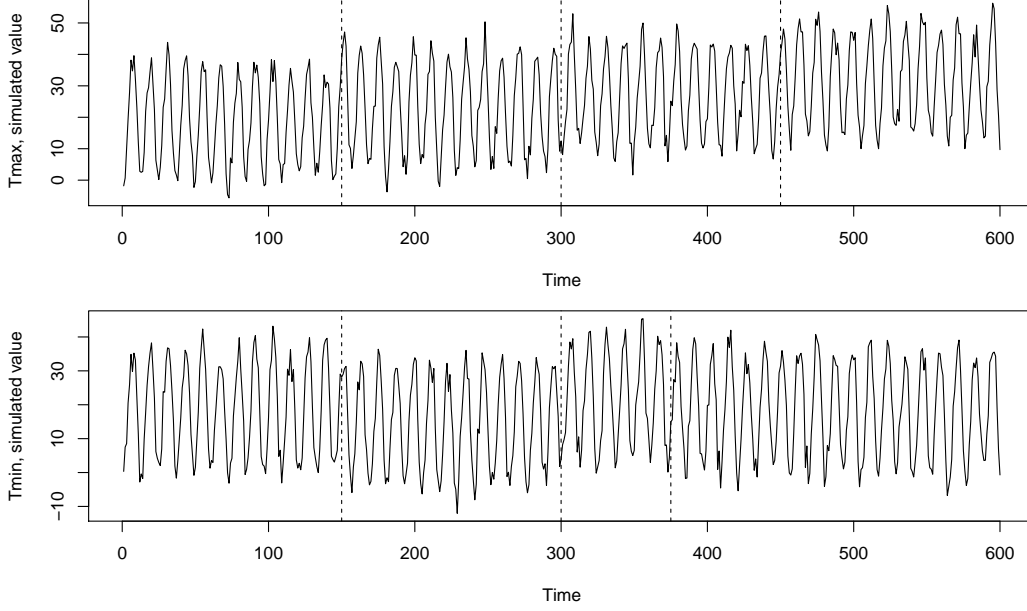


Figure 1: A simulated dataset with the signal to noise ratio $\kappa = 1.5$, which has three change-points in Tmax (top panel) and three changepoints in Tmin (bottom panel). Vertical dashed lines mark true changepoint times.

subtraction of sample monthly means.

5.1 Univariate simulations

First, the Tmax and Tmin series are analyzed separately, each fitted by univariate BMDL methods with default parameters, once with the fictitious metadata and once without metadata. We also compare various methods without metadata, including BMDL under the objective Bayes parameters $a = b = 1$ (denoted by oBMDL), the automatic MDL (15) (denoted by MDL), and the BIC, where $\text{BIC}(\boldsymbol{\eta}) = (N - p) \log(\hat{\sigma}_{\nu=\infty}^2) / 2 + m \log(N - p)$ up to a constant. In each fit, a MCMC chain of 100,000 iterations is generated. The optimal multiple changepoint model is taken as the one minimizes the objective function.

For Tmax series, Table 1 reports empirical detection percentages, including true positive rates at the exact times of changepoints and average false positive rates at non-changepoint times, along with estimated number of changepoints \hat{m} and its standard error. When metadata is ignored, since the three shifts are of equal size Δ , their detection rates are similar.

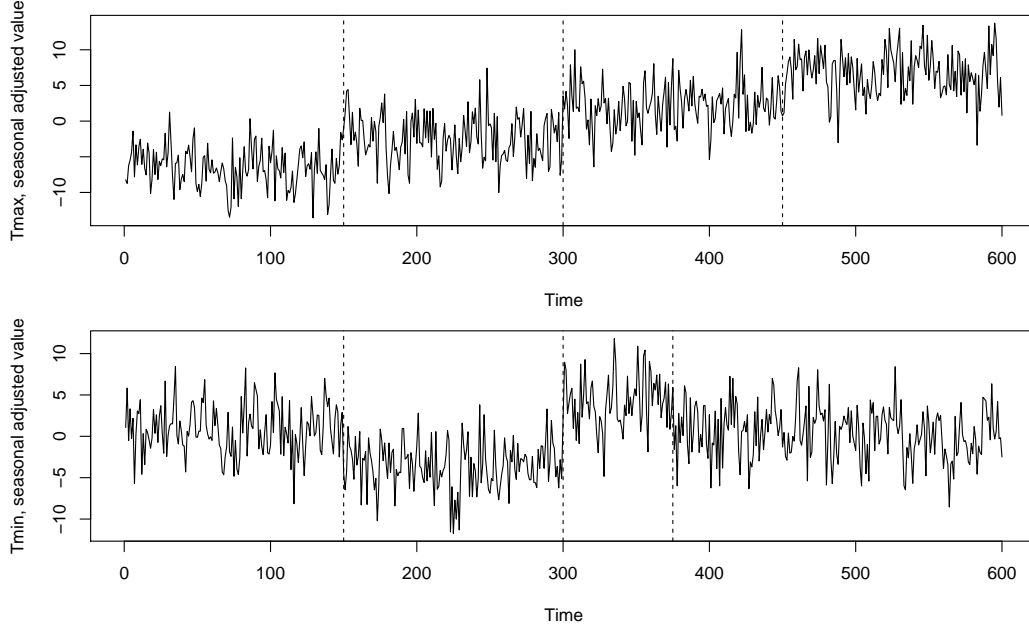


Figure 2: The Figure 1 series after subtracting sample monthly means. Vertical dashed lines mark true changepoint times.

False detection rates are very low; even when $\kappa = 1$, on average, a non-changepoint is detected 0.43% of the time or less.

Among different methods without metadata, detection rates of true changepoints are similar, while BIC flags slightly more false positives than MDL-based methods (BMDL, oBMDL, and MDL). When $\kappa = 1$, the number of changepoints $m = 3$ is underestimated by the MDL-based methods and better estimated by BIC penalties; when $\kappa = 1.5$ and 2, m is correctly estimated by the MDL-based methods, and overestimated by BIC. Overall, BIC tends to favor models with more changepoints than the MDL-based methods. As suggested by Proposition 1, the BMDL performs similarly to the automatic MDL.

Interestingly, without metadata, the BMDL under the default parameters $a = 1$ and $b = 239$ and the objective choices $a = b = 1$ perform similarly. Figure 3 reveals that as functions of m , the code lengths $\mathcal{L}(\boldsymbol{\eta})$ under BMDL and oBMDL have similar shapes, with a nearly constant difference over the region where m is small. Therefore, if knowledge of changepoint frequency is not available, a BMDL can still be used with objective parameters.

Metadata use substantially increases detection power for the BMDL. In Figure 4, the true

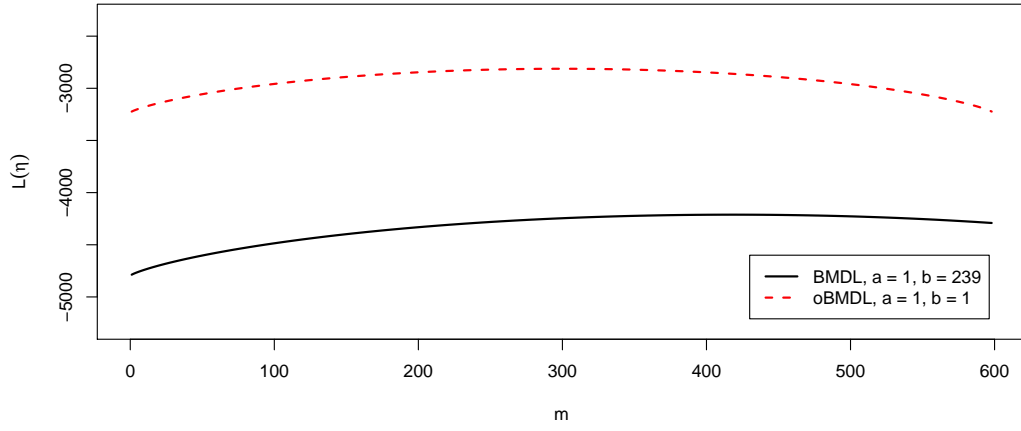


Figure 3: Model code lengths $\mathcal{L}(\boldsymbol{\eta}) = -\log \Gamma(a+m) - \log \Gamma(b+N-p-m)$ between the BMDL and the oBMDL.

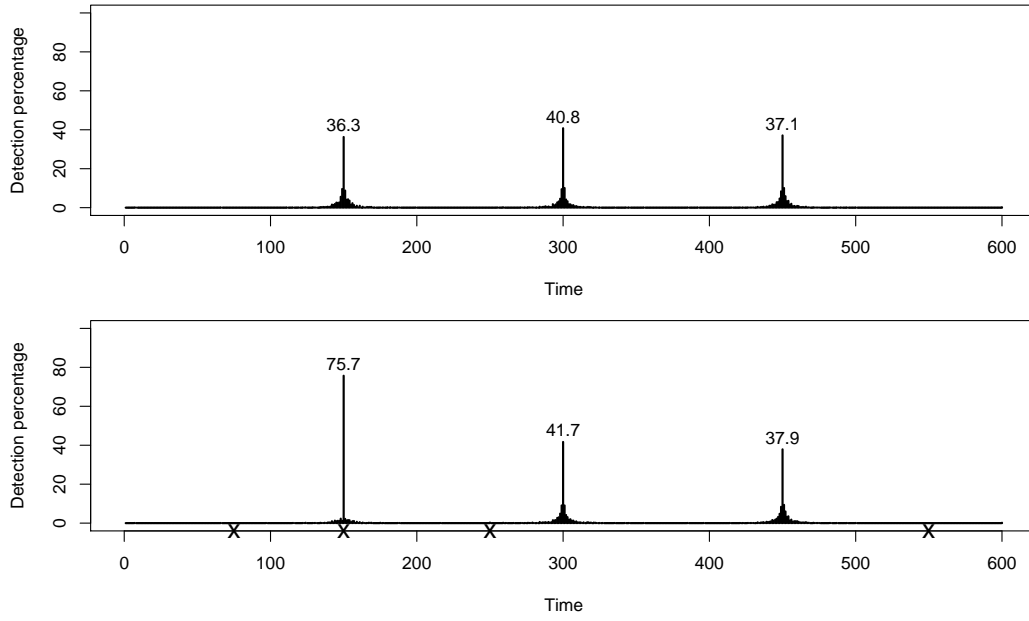


Figure 4: Detection times and percentages of changepoints in Tmax series using univariate BMDL. The top panel ignores the four metadata times; the bottom panel uses the metadata (metadata times are marked as crosses on the axis). Numerical percentages on the graphic are for detection at “their exact times”. The results are aggregated from 1000 independent simulated datasets with $\kappa = 1.5$.

Table 1: Univariate results for Tmax, aggregated from 1000 simulated datasets. The detection rates for the documented change when metadata is used are in bold.

κ	Metadata	Method	True positive detection (%)			Average false positive detection (%)	\hat{m} (se)
			$t = 150$	$t = 300$	$t = 450$		
1.0	yes	BMDL	58.8	16.8	14.5	0.29	2.65 (0.56)
	no	BMDL	15.4	16.3	16.4	0.36	2.61 (0.61)
	no	oBMDL	14.4	16.9	16.1	0.37	2.68 (0.59)
	no	MDL	14.9	17.2	16.2	0.36	2.64 (0.62)
	no	BIC	17.0	17.4	18.3	0.43	3.07 (0.54)
1.5	yes	BMDL	75.7	41.7	37.9	0.25	3.02 (0.13)
	no	BMDL	36.3	40.8	37.1	0.31	3.02 (0.13)
	no	oBMDL	36.5	41.3	37.2	0.31	3.03 (0.17)
	no	MDL	37.6	41.3	37.0	0.31	3.02 (0.15)
	no	BIC	37.0	40.2	36.3	0.33	3.12 (0.38)
2.0	yes	BMDL	84.1	59.3	57.6	0.17	3.02 (0.14)
	no	BMDL	54.2	59.7	57.2	0.22	3.02 (0.15)
	no	oBMDL	54.4	59.4	57.3	0.22	3.03 (0.18)
	no	MDL	54.7	59.4	58.0	0.22	3.02 (0.16)
	no	BIC	53.4	59.1	56.9	0.24	3.11 (0.36)

documented change at time 150 is detected 75.7% of the time when metadata is used, more than twice as high (36.3%) when metadata is eschewed. Moreover, times near the changepoint at time 150 are less likely to be flagged as changepoints. Our prior belief that metadata times are more likely to be changepoints is especially important when the mean shift is small: when $\kappa = 1$, using metadata increases the detection rate of the time 150 changepoint from 15.4% to 58.8%. For false positives, Figure 4 shows that using metadata does not increase false detection rates at the documented times 75, 250, and 550 (where no shifts occur). This suggests that the prior distribution does not “overwhelm” the data. Table 1 shows that average false positive rates even drop after using metadata.

For Tmin series, the non-monotonic shift aspect (down, up, down) that troubles at most one change (AMOC) binary segmentation approaches (Li and Lund 2012) is well handled by all multiple changepoint detection methods examined. Table 2 shows that when metadata is ignored, the larger shift at time 300 is more easily detected than the two smaller shifts at times 150 and 375. When metadata is used, the detection rate of the time 150 shift becomes

Table 2: Univariate results for Tmin, aggregated from 1000 simulated datasets. Detection rates for the documented change when metadata is used are in bold.

κ	Metadata	Method	True positive detection (%)			Average false positive detection (%)	\hat{m} (se)
			$t = 150$	$t = 300$	$t = 375$		
1.0	yes	BMDL	62.0	53.5	14.3	0.23	2.69 (0.77)
	no	BMDL	18.0	52.4	14.1	0.30	2.63 (0.86)
	no	oBMDL	18.7	54.9	14.6	0.31	2.76 (0.71)
	no	MDL	17.4	50.5	13.6	0.28	2.50 (0.99)
	no	BIC	19.5	55.0	15.8	0.36	3.07 (0.52)
1.5	yes	BMDL	77.3	84.4	38.2	0.17	3.01 (0.15)
	no	BMDL	37.4	84.7	39.5	0.24	3.02 (0.17)
	no	oBMDL	37.5	84.3	38.9	0.24	3.03 (0.20)
	no	MDL	37.2	84.3	38.6	0.24	3.01 (0.15)
	no	BIC	36.5	83.3	38.0	0.26	3.13 (0.44)
2.0	yes	BMDL	85.2	95.4	56.1	0.11	3.01 (0.13)
	no	BMDL	58.2	95.4	56.4	0.15	3.02 (0.13)
	no	oBMDL	58.2	95.2	56.5	0.16	3.03 (0.18)
	no	MDL	58.0	95.5	56.9	0.15	3.01 (0.12)
	no	BIC	57.7	95.5	55.7	0.17	3.12 (0.43)

comparable to the detection rate of time 300 shift, which is twice as large in size, but is not a metadata time. False positive rates are uniformly low, and m is well-estimated by MDL-based methods when κ is not too small. Again, without metadata, the MDL-based methods are similar, while BIC tends to favor models with larger m .

5.2 Bivariate simulations

Since the BMDL is flexible enough to handle non-concurrent shifts for bivariate series, we now apply it to Tmax and Tmin series jointly. Each bivariate series is fitted by a MCMC chain of 50,000 iterations, once without metadata, and once with metadata. Metadata impacts are similar to the univariate case, increasing detection of true mean shifts at metadata times and also slightly decreasing average false positive rates (see Tables 3 and 4). Figure 5 shows bivariate detection rates with metadata when $\kappa = 1.5$. For the non-concurrent shifts times 375 and 450, detection rates for the component series are very different; in most runs, concurrent shifts are not flagged.

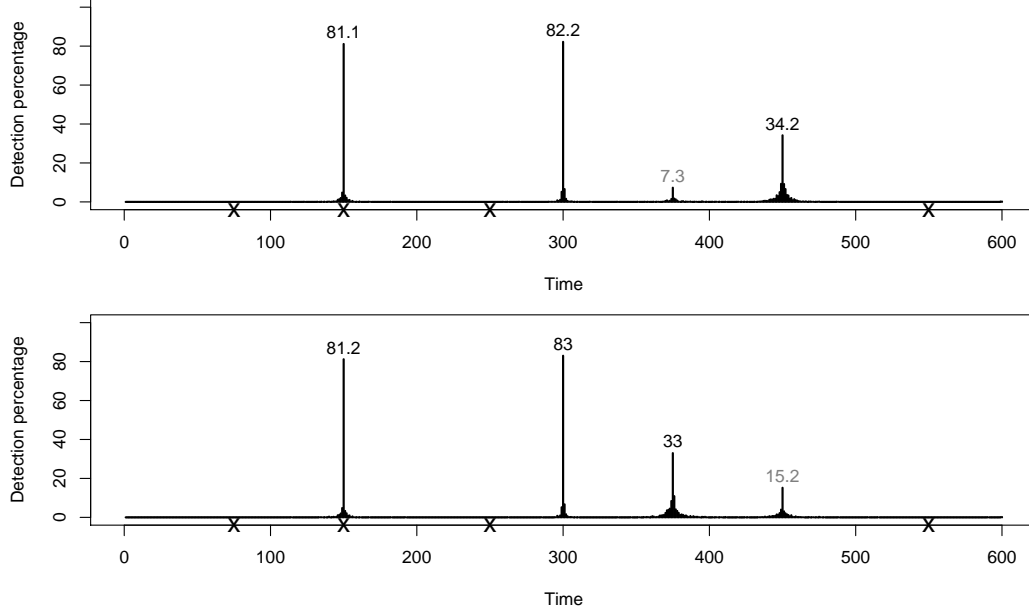


Figure 5: Detection percentages of Tmax (top panel) and Tmin (bottom panel) using bivariate BMDL methods with metadata (metadata times are marked as crosses on the axis). Numerical percentages on the graphic are for detection at “their exact times”. The results are aggregated from 1000 independent simulated datasets with $\kappa = 1.5$.

While concurrent shifts are not always the case, they are believed to be more likely in our parameter elicitation settings. Compared to the univariate BMDL, the bivariate method enhances detection power of concurrent changepoints. When $\kappa = 1.5$, at time 150, where Tmax (Tmin) shifts Δ ($-\Delta$), the bivariate BMDL increases the univariate detection rates from both series from about 77% to above 81%. At time 300, where Tmax (Tmin) shifts by Δ (2Δ), the detection rate increases from 41.1% to 82.2% for Tmax, while it slightly drops for Tmin. Tables 3 and 4 show that detection power gains under the bivariate approach are greater for small signals $\kappa = 1$, without metadata. An interesting phenomenon is observed: bivariate BMDL improves univariate methods more when the concurrent shifts move the series in opposite directions than in the same direction (detection rates at time 300 do not increase for Tmin). Because Tmax and Tmin are positively correlated series, concurrent shifts in the same direction may be misattributed to positively correlated errors; this cannot happen for shifts in opposite directions.

Table 3: Bivariate results for Tmax by BMDL, aggregated from 1000 simulated datasets.

κ	Metadata	True positive detection (%)			False positive detection (%)		\hat{m} (se)
		$t = 150$	$t = 300$	$t = 450$	$t = 375$	average	
1.0	yes	60.7	54.5	11.5	6.8	0.31	3.12 (0.45)
	no	36.5	55.2	11.4	8.3	0.36	3.19 (0.48)
1.5	yes	81.1	82.2	34.2	7.3	0.20	3.18 (0.43)
	no	66.7	82.9	33.9	10.8	0.24	3.29 (0.47)
2.0	yes	92.1	93.5	55.9	3.7	0.11	3.07 (0.28)
	no	84.7	94.8	55.6	6.2	0.13	3.13 (0.35)

Table 4: Bivariate results for Tmin by BMDL, aggregated from 1000 simulated datasets.

κ	Metadata	True positive detection (%)			False positive detection (%)		\hat{m} (se)
		$t = 150$	$t = 300$	$t = 375$	$t = 450$	average	
1.0	yes	60.1	54.9	9.5	8.7	0.31	3.10 (0.57)
	no	36.2	55.3	10.2	9.6	0.36	3.17 (0.55)
1.5	yes	81.2	83.0	33.0	15.2	0.24	3.38 (0.54)
	no	66.4	83.4	34.2	21.3	0.30	3.61 (0.54)
2.0	yes	92.0	94.8	57.8	16.2	0.14	3.28 (0.46)
	no	84.8	95.1	54.9	32.1	0.21	3.59 (0.53)

Overall, while bivariate detection does not induce more false positives, it tends to flag more false positives at locations where the mean in the other series shifts. Figure 5 shows that at time 375, a changepoint time in Tmin but not in Tmax, a false detection rate of 7.3% for Tmax is obtained. At time 450, a changepoint in Tmax but not Tmin, a false detection rate of 15.2% is obtained for Tmin. These false positive rates slightly degrade inferences at nearby changepoints; for example, at time 450 for Tmax and time 375 for Tmin, detection rates are 34.2% and 33.0%, respectively, slightly lower than the 37.9% and 38.2% reported in the univariate case. Finally, Tables 3 and 4 show that the bivariate approach tends to overestimate m , which differs from the univariate case.

6 The Tuscaloosa Data

A monthly Tmax and Tmin series from Tuscaloosa, Alabama (the target station) over the 114 year period January, 1901 – December, 2014 is plotted in Figure 7. Lu et al. (2010) study

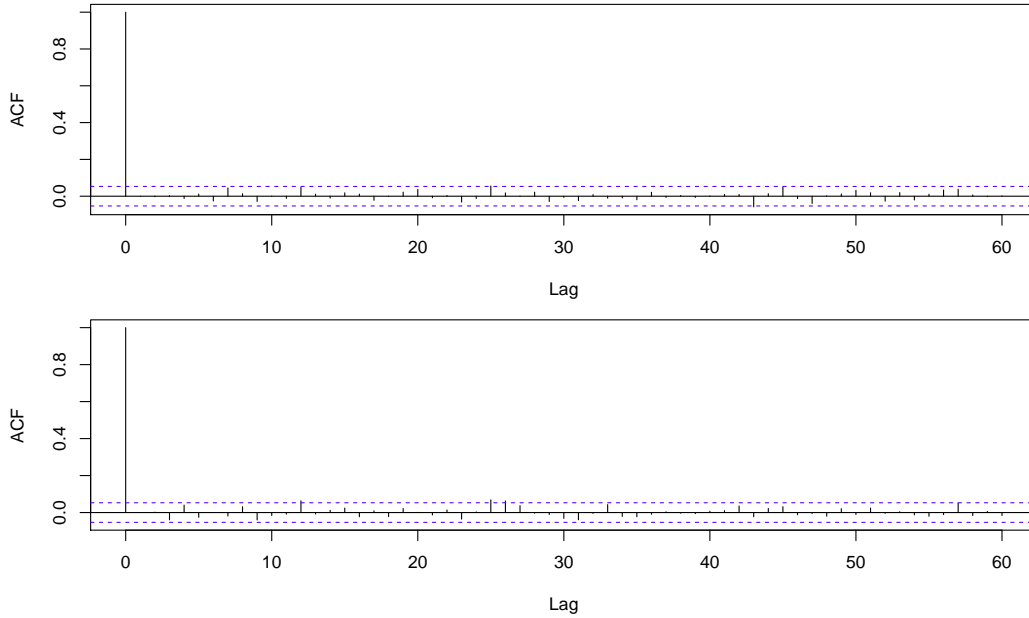


Figure 6: Sample model residual autocorrelations for Tmax (top panel) and Tmin (bottom panel), fitted using the univariate BMDL with metadata and $p = 2$.

annually averaged values of this series from 1901-2000. The Tuscaloosa metadata lists station relocations in November 1921, March 1939, June 1956, and May 1987; November 1956 and May 1987 are listed as instrumentation change times.

In this section, the Tmax and Tmin series will be analyzed from both univariate and bivariate perspectives via the penalization methods examined in Section 5. All parameters are set to default values; the AR order $p = 2$ is judged as appropriate: in Figure 6, almost all sample autocorrelations of residuals fitted with $p = 2$ lie inside the pointwise 95% confidence bands.

To ensure convergence in the MCMC search algorithm, for each fit, 50 Markov chains are generated from different starting points, each containing 1,000,000 (univariate) or 100,000 (bivariate) iterations. Among all changepoint models visited by the 50 Markov chains, the one with the smallest BMDL is reported as the optimal model.

6.1 Univariate fits

The top half of Table 5 displays estimated changepoints for the univariate fits. When metadata is ignored, all methods (BMDL, oBMDL, MDL, and BIC) estimate the same optimal changepoint configuration: Tmax has two estimated changepoints and Tmin has three; of these, only January 1990 is a concurrent change. Another changepoint is approximately concurrent: March 1957 for Tmax and July 1957 for Tmin. The 1918 changepoint flagged for Tmin is close to the station relocation in November 1921; the station relocation in June 1956 and the equipment change in November 1956 are near the two estimated changepoints in 1957. The metadata time in May 1987 is about three years from the concurrent changepoints flagged in January 1990. Of course, when metadata is ignored, estimated changepoint times may not coincide (exactly) with metadata times.

Table 5: Estimated changepoints for the Tuscaloosa data.

Metadata	Series	Estimated changepoints
Univariate		
yes	Tmax	1956 Nov, 1987 May
	Tmin	1921 Nov, 1956 Jun, 1987 May
no	Tmax	1957 Mar, 1990 Jan
	Tmin	1918 Feb, 1957 Jul, 1990 Jan
Bivariate		
yes	Tmax	1921 Nov, 1956 Jun, 1987 May
	Tmin	1921 Nov, 1956 Jun, 1987 May
no	Tmax	1918 Feb, 1957 Jul, 1988 Jul
	Tmin	1918 Feb, 1957 Jul, 1988 Jul

Repeating the above analysis with metadata, two changepoints are found in Tmax and three in Tmin. All estimated changepoint times now coincide with metadata times. Only the May 1987 changepoint is concurrent. Between Tmax and Tmin, the two estimated changepoints in 1956 (i.e., the two metadata times in 1956) are just a few months apart. As parameter estimates are similar with or without metadata, only estimates for the optimal changepoint model with metadata are reported. For Tmax, estimated regime means are (standard errors in parentheses) $\hat{\mu}_2 = -1.50$ (0.24) and $\hat{\mu}_3 = 0.66$ (0.25) (recall that $\mu_1 = 0$); estimated AR(2)

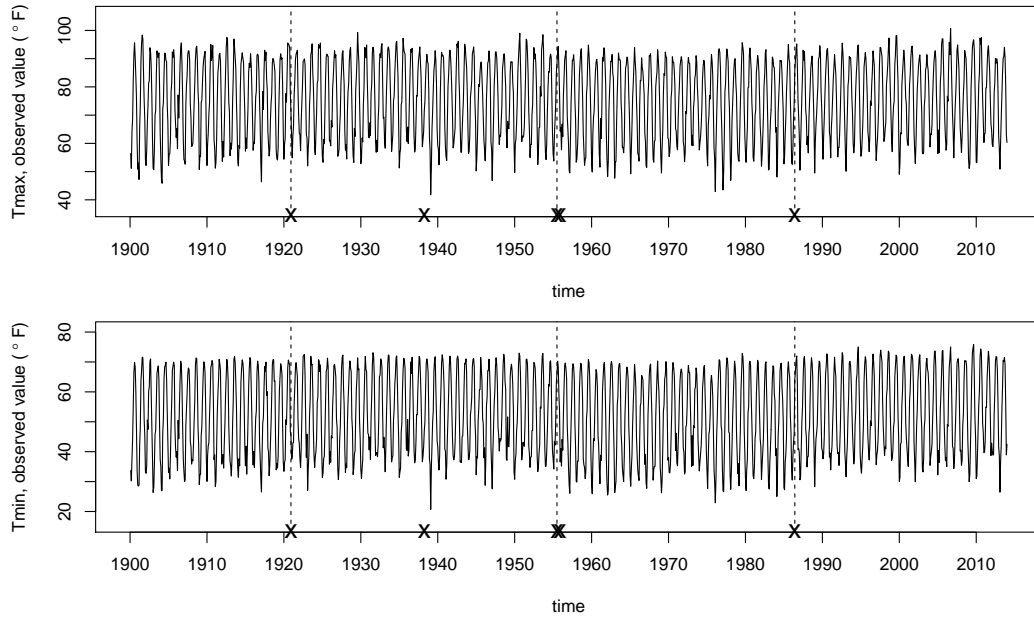


Figure 7: Tuscaloosa monthly Tmax (top panel) and Tmin (bottom panel) series. Metadata times are marked with crosses on the axis. Vertical dashed lines show estimated changepoint times from bivariate BMDL with metadata.

coefficients are $\hat{\phi}_1 = 0.21$, $\hat{\phi}_2 = 0.05$, and $\hat{\sigma}^2 = 11.59$. For Tmin, the estimated parameters are $\hat{\mu}_2 = 1.76$ (0.21), $\hat{\mu}_3 = -1.06$ (0.22), $\hat{\mu}_4 = 2.35$ (0.24), $\hat{\phi}_1 = 0.18$, $\hat{\phi}_2 = 0.05$, and $\hat{\sigma}^2 = 10.81$. The concurrent May 1987 changepoint shifts both series to warmer regimes.

6.2 Bivariate fits

Both Tmax and Tmin series are now analyzed in tandem by BMDL. Three changepoints are detected in both series, with or without metadata, and all are concurrent (see the bottom half of Table 5). Figure 7 illustrates the optimal bivariate BMDL changepoint configuration. When metadata is used, all estimated changepoint times migrate to metadata times. Comparing to the univariate results, the bivariate approach yields the same changepoint configuration for Tmin; for Tmax, a new changepoint in November 1921 is flagged and the November 1956 changepoint moves to June 1956, thus becoming a concurrent change. For this changepoint

configuration, the estimated VAR parameters are

$$\hat{\Phi}_1 = \begin{pmatrix} 0.21 & -0.01 \\ -0.02 & 0.20 \end{pmatrix}, \quad \hat{\Phi}_2 = \begin{pmatrix} 0.06 & -0.02 \\ -0.04 & 0.08 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 11.56 & 8.13 \\ 8.13 & 10.81 \end{pmatrix}.$$

In temperature homogenization problems, the goal is often to detect (and then adjust for) “artificial” changes. Naturally occurring climate shifts should be left in the record if possible. Because of this, analyses often consider target minus reference series. A reference series is a record from a station near the target station that is subtracted from the target series. The idea is that two nearby stations should experience similar weather; hence, any trends or seasonal cycles should be lessened (if not altogether removed) in the target minus reference subtraction. Changepoints are easier to detect (visually) in target minus reference comparisons. Following [Lu et al. \(2010\)](#), our reference series is obtained by averaging three nearby stations: Aberdeen, MS; Greensboro, AL; and Selma, AL. By averaging multiple reference series (this is called a composite reference), impacts of mean shifts in any of the individual stations in the composite reference are minimized.

Figure 8 shows the optimal changepoint configuration for the target minus reference series and contains 12 concurrent changes: June 1914, January 1919, July 1933, July 1937, August 1937, October 1938, December 1938, June 1946, July 1946, November 1956, May 1987, and October 1996. Among them, the 1956 and 1987 changepoints are in the metadata; the two changepoints in 1938 are close to the 1939 station relocation. The changepoints in 1919, 1933, and 1990 are also flagged by [Lu et al. \(2010\)](#). One of the shifts, November 1956, moves the Tmax series warmer and the Tmin series colder.

The October and December 1938 changepoints are likely due to typos in the data record. Specifically, the October and November 1938 Tmin values in the target minus reference series appear to be abnormally high. While the data have been quality checked, some errors persist. This conjecture is made because the three reference stations lie in various directions from Tuscaloosa; climatologically, series to the north and west of Tuscaloosa should be cooler and

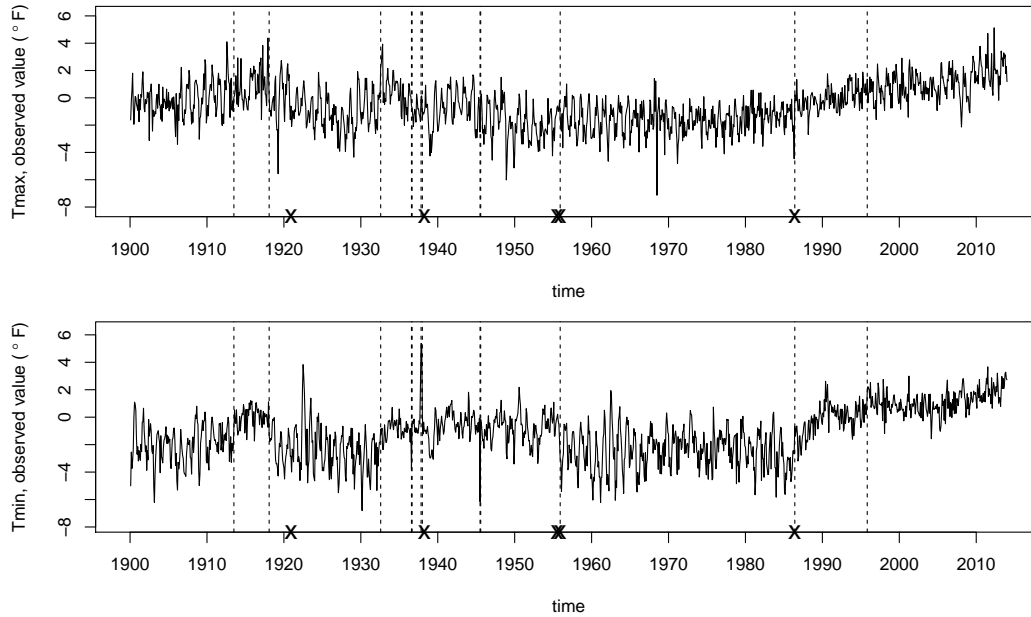


Figure 8: Target minus reference Tmax (top panel) and Tmin (bottom panel) series. Metadata times for Tuscaloosa are marked with crosses on the axis. Vertical dashed lines show estimated changepoint times from our methods.

those to the south and east should be warmer. In this case, Tuscaloosa was significantly warmer than all three references. Similar statements apply to the two “outlier” changepoints in 1937, and the two changepoints in 1946, where the Tmin records for Tuscaloosa are lower than those for all three reference stations. It is interesting that our method picks up outliers.

It is natural to flag more changepoints in the target minus reference series than the target series alone. An ideal reference series should have the same trend and seasonal cycles as the target series and be free of artificial mean shifts. This said, we do not assume that the target minus reference comparison completely removes the monthly mean cycle; indeed, [Liu et al. \(2015\)](#) shows that this is seldom the case. Reference series selection is a problem currently studied by climatologists. As our reference series averages three neighbor stations, mean shifts in any of the reference records may induce shifts in the target minus reference series. For example, the estimated changepoint in 1914 is close to the 1915 metadata time listed in the Aberdeen reference. This said, averaging three neighbors should help mitigate the effects of changepoints in any individual reference series.

7 Discussion

This paper develops a flexible MDL-based multiple changepoint detection approach to accommodate various prior distributional specifications. Motivated by climate homogenization problems, our Bayesian MDL method incorporates subjective knowledge such as metadata in mean shift detection for univariate autoregressive processes with seasonal means, and then extended these ideas to bivariate VAR settings. Both theoretical and simulation studies show that without metadata, our BMDL performs similarly to the state-of-art automatic MDL method; with metadata, the BMDL’s detection power significantly improves. Our BMDL has several practical advantages, including simple parameter elicitation, asymptotic consistency, and efficient MCMC computation.

The approach can be extended to accommodate more flexible time series structures, including moving-averages, periodic autoregressions, and more than two series. The methods could also be tailored to categorical data. For count data, the likelihood could be Poisson-based. With a conjugate Gamma prior, the resulting marginal likelihoods will again have closed forms. There is no technical difficulty in allowing a background linear trend, or even piecewise linear trends. This said, linear trends can be mistaken for multiple mean shifts should trends be present and ignored in the analysis ([Li and Lund 2015](#)).

Non-MCMC stochastic search methods could also be used. Genetic algorithms, popular in multiple changepoint MDL analyses, are also capable of minimizing the BMDL. Pre-screening methods such as [Chan et al. \(2014\)](#); [Yau and Zhao \(2015\)](#) can speed up model search algorithms. When no global parameters exist in the likelihood (i.e., independent observations, no seasonal cycle, error variance known), dynamic programming based techniques such as the PELT ([Killick et al. 2012](#)) can further accelerate computational speed.

Supplementary Material

Appendix: Proof of Theorem 1 This supplement provides an asymptotic analysis of the

Yule-Walker estimator $\hat{\phi}$ under any candidate relative changepoint model λ . A proof of the asymptotic model selection consistency in Theorem 1 is also given.

Acknowledgement

The authors thank Matthew Menne, Jared Rennie, Claude Williams Jr., and Bin Yu for helpful discussions. The climate application was posed at SAMSI's 2014 climate homogeneity summit in Boulder, Colorado. Robert Lund and Hewa A. Priyadarshani thank NSF Grant DMS 1407480 for partial support. Clemson University is acknowledged for generous allotment of computation time on its Palmetto cluster.

Supplementary Material for “Bayesian Minimal Description Lengths for Multiple Changepoint Detection.”

A Proof of Theorem 1

A.1 Asymptotic behavior of the Yule-Walker estimator $\hat{\phi}$

To prove Theorem 1, the asymptotic limit of the Yule-Walker estimator is needed. For a sample size N , the observations obey the true changepoint model $\boldsymbol{\lambda}^0$ in (5):

$$\mathbf{X} = \mathbf{A}\mathbf{s} + \mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}.$$

Here, $\boldsymbol{\epsilon}$ is a zero-mean causal AR(p) series. For notation, the symbols $\mathbf{s}, \sigma^2, \boldsymbol{\phi}$ refer to the true parameters in $\boldsymbol{\lambda}^0$. Moreover, the subscript $1 : N$ is omitted wherever there is no ambiguity.

For any relative changepoint model $\boldsymbol{\lambda}$, suppose that $\boldsymbol{\eta}$ is the corresponding changepoint configuration under the sample size N . From (12), the ordinary least squares residual vector

$$\boldsymbol{\epsilon}^{\text{ols}} = (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]})\mathbf{X} = (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]})(\mathbf{A}\mathbf{s} + \mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}) = (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]})(\mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}). \quad (16)$$

Here, the regime matrix \mathbf{D} depends on $\boldsymbol{\eta}$ and may not equal \mathbf{D}^0 .

Lemma 1. *For each relative changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and $t = 1, \dots, N$, when N is large, each entry of $\boldsymbol{\epsilon}^{\text{ols}}$ can be expressed as $\epsilon_t^{\text{ols}} = \delta_t + W_t$, where*

$$\delta_t = \mu_{r^0(t)}^0 - \bar{\mu}_{r(t)}^0, \quad W_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}.$$

Here, in regime ℓ of the changepoint configuration $\boldsymbol{\lambda}$, $\bar{\mu}_\ell^0 = (N_\ell)^{-1} \sum_{t \in \mathcal{R}_\ell} \mu_t^0$ is the average of the true mean parameters, N_ℓ is the number of time points in this regime, and \mathcal{R}_ℓ is the set of all time points in this regime. Likewise, $\bar{\epsilon}_\ell$ is the average of errors in regime ℓ , $\bar{\epsilon}_v$ is the

average of errors in season v , and $\bar{\epsilon}$ is the average of all errors.

Proof. Because of (16), our main effort is to study the projection residual $\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]}$ under large N . Since the two column spaces spanned by $(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}$ and \mathbf{D} are perpendicular, Theorem B.45 in Christensen (2002, pp. 411) gives $\mathcal{P}_{[(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A} \ \mathbf{D}]} = \mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} + \mathcal{P}_{\mathbf{D}}$. Therefore,

$$\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]} = \mathbf{I}_N - \mathcal{P}_{[(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A} \ \mathbf{D}]} = \mathbf{I}_N - \mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} - \mathcal{P}_{\mathbf{D}}. \quad (17)$$

Here, the term $\mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}}$ is expanded as

$$\mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} = (\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A} [\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}]^{-1} \mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}}). \quad (18)$$

For any $n \in \mathbb{N}$, let $\mathbf{0}_n$ be the n -dimensional vector containing all zero entries, $\mathbf{1}_n$ be the n -dimensional vector whose entries are all unity, and \mathbf{J}_n as the $n \times n$ matrix whose entries are all unity, i.e., $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n'$.

For $v = 1, \dots, T$, suppose there are $k(v, \ell)$ time points in regime ℓ that are also in season v . Since N_ℓ increases linearly with N , so does $k(v, \ell)$. Moreover, when N is large, inside each regime, the seasonal counts $k(v, \ell)$ are equal except for edge effects, i.e., $k(v, \ell)/N_\ell \approx 1/T$ for all seasons v . We will ignore these edge effects in the ensuing calculations. Proceeding under this simplification, the v th column in \mathbf{A} , denoted by \mathbf{A}_v , under the projection $\mathcal{P}_{\mathbf{D}}$, becomes

$$\mathcal{P}_{\mathbf{D}}\mathbf{A}_v = \left(\mathbf{0}'_{N_1}, \frac{k(v, 2)}{N_2} \mathbf{1}'_{N_2}, \dots, \frac{k(v, m+1)}{N_{m+1}} \mathbf{1}'_{N_{m+1}} \right)' = \left(\mathbf{0}'_{N_1}, \frac{1}{T} \mathbf{1}'_{N-N_1} \right)'. \quad (19)$$

We can now obtain an expression for $\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}$. To do this, note that for $u, w \in \{1, 2, \dots, T\}$,

$$[\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}]_{u,w} = \mathbf{A}'_u \mathbf{A}_w - (\mathcal{P}_{\mathbf{D}}\mathbf{A}_u)'(\mathcal{P}_{\mathbf{D}}\mathbf{A}_w) = \begin{cases} \frac{N}{T^2}(T - (1 - \lambda_1)), & \text{if } u = w, \\ -\frac{N}{T^2}(1 - \lambda_1), & \text{if } u \neq w, \end{cases}$$

and it follows that $\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A} = NT^{-2}(T\mathbf{I}_T - (1 - \lambda_1)\mathbf{J}_T)$. The inverse of this matrix can be verified as

$$[\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}]^{-1} = \frac{1}{N} \left(T\mathbf{I}_T + \frac{1 - \lambda_1}{\lambda_1} \mathbf{J}_T \right).$$

Plugging this inverse into (18) and denoting $\mathcal{Q}_{\mathbf{D}} = \mathbf{I}_N - \mathcal{P}_{\mathbf{D}}$ give

$$\begin{aligned} \mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} &= \frac{1}{N} (\mathcal{Q}_{\mathbf{D}}\mathbf{A}) \left(T\mathbf{I}_T + \frac{1 - \lambda_1}{\lambda_1} \mathbf{J}_T \right) (\mathcal{Q}_{\mathbf{D}}\mathbf{A})' \\ &= \frac{T}{N} (\mathcal{Q}_{\mathbf{D}}\mathbf{A})(\mathcal{Q}_{\mathbf{D}}\mathbf{A})' + \frac{1 - \lambda_1}{N\lambda_1} (\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)'. \end{aligned} \quad (20)$$

For simplicity, we assume that regime ℓ starts with season 1, ends with season T , and contains n_ℓ full cycles. Using $n = N/T = \sum_{r=1}^{m+1} n_r$ and (19) gives

$$\mathcal{Q}_{\mathbf{D}}\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \mathbf{I}_T \\ \hline \mathbf{1}_{n-n_1} \otimes (\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T) \end{pmatrix}, \quad \mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T = \begin{pmatrix} \mathbf{1}_{N_1} \\ \hline \mathbf{0}_{N-N_1} \end{pmatrix}.$$

Hence, quadratic forms of these matrices are

$$(\mathcal{Q}_{\mathbf{D}}\mathbf{A})(\mathcal{Q}_{\mathbf{D}}\mathbf{A})' = \begin{pmatrix} \mathbf{J}_{n_1} \otimes \mathbf{I}_T & \vdots & \mathbf{J}_{n_1 \times (n-n_1)} \otimes (\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T) \\ \hline \mathbf{J}_{(n-n_1) \times n_1} \otimes (\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T) & \vdots & \mathbf{J}_{n-n_1} \otimes (\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T) \end{pmatrix},$$

and

$$(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)' = \begin{pmatrix} \mathbf{J}_{N_1} & \vdots & \mathbf{0} \\ \hline \mathbf{0} & \vdots & \mathbf{0} \end{pmatrix}.$$

Plugging these into (20) produces

$$\mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} = \frac{1}{N_1} \begin{pmatrix} \mathbf{J}_{N_1} & \vdots & \mathbf{0} \\ \hline \mathbf{0} & \vdots & \mathbf{0} \end{pmatrix} + \frac{T}{N} \mathbf{J}_n \otimes \mathbf{I}_T - \frac{1}{N} \mathbf{J}_N.$$

Since $\mathcal{P}_{\mathbf{D}}$ is block-diagonal of form

$$\mathcal{P}_{\mathbf{D}} = \text{diag} \left(\mathbf{0}_{N_1 \times N_1}, \frac{\mathbf{J}_{N_2}}{N_2}, \dots, \frac{\mathbf{J}_{N_{m+1}}}{N_{m+1}} \right),$$

we have

$$\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]} = \mathbf{I}_N - \text{diag} \left(\frac{\mathbf{J}_{N_1}}{N_1}, \frac{\mathbf{J}_{N_2}}{N_2}, \dots, \frac{\mathbf{J}_{N_{m+1}}}{N_{m+1}} \right) - \frac{T}{N} \mathbf{J}_n \otimes \mathbf{I}_T + \frac{1}{N} \mathbf{J}_N.$$

Therefore, for $t \in \{1, 2, \dots, N\}$, the t th entries of the vectors in (16) are

$$\begin{aligned} W_t &= [(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]})\boldsymbol{\epsilon}]_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}, \\ \delta_t &= [(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A} \ \mathbf{D}]})\mathbf{D}^0 \boldsymbol{\mu}^0]_t = \mu_{r^0(t)}^0 - \bar{\mu}_{r(t)}^0. \end{aligned}$$

□

It is not hard to see that $\delta_t = 0$ for all $t = 1, \dots, N$ if and only if all relative changepoints in $\boldsymbol{\lambda}^0$ are contained in $\boldsymbol{\lambda}$ (denoted by $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$). For any changepoint configuration $\boldsymbol{\lambda}$, as N tends to infinity, the average $N^{-1} \sum_{t=h+1}^N \delta_t \delta_{t-h}$ converges to a constant that does not depend on the lag $h = 0, 1, \dots, p$. This is because for any lag h , $\delta_t = \delta_{t-h}$ for all $t = 1, \dots, N$, except for at most $(m + m^0)h \leq (m + m^0)p$ times near the changepoints in $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^0$. Hence, as $N \rightarrow \infty$, $N^{-1} \sum_{t=h+1}^N \delta_t \delta_{t-h}$ converges to its limit at rate $O(1/N)$. We denote this limit as

$$\delta^2 \stackrel{\text{def}}{=} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=h+1}^N \delta_t \delta_{t-h}. \quad (21)$$

Since (21) holds for $h = 0$, $\delta^2 \geq 0$.

To quantify the asymptotic limit of the Yule-Walker estimator $\hat{\boldsymbol{\phi}}$, let $\boldsymbol{\gamma}_p = (\gamma(1), \dots, \gamma(p))'$ and $\boldsymbol{\Gamma}_p$ be a $p \times p$ matrix with (i, j) th entry $\gamma(|i - j|)$.

Proposition 2. *Under the relative changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ (which may or may*

not be the true changepoint configuration), for $h = 0, 1, \dots, p$, as $N \rightarrow \infty$, the lag h sample autocovariance obeys

$$\hat{\gamma}(h) = \gamma(h) + \delta^2 + O_P\left(\frac{1}{\sqrt{N}}\right),$$

and the Yule-Walker estimator $\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p$ obeys

$$\hat{\phi} = (\Gamma_p + \delta^2 \mathbf{J}_p)^{-1} (\gamma_p + \delta^2 \mathbf{1}_p) + O_P\left(\frac{1}{\sqrt{N}}\right). \quad (22)$$

Moreover, if and only if $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$, $\delta^2 = 0$ and $\hat{\phi} \rightarrow \phi$ as $N \rightarrow \infty$.

Proof. Since the AR(p) errors are assumed causal, we may write

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

for some weights $\{\psi_j\}_{j=0}^{\infty}$, where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Since $W_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}$, one can write W_t as a linear combination of all Z_t s up to and before time N :

$$W_t = \sum_{j=-\infty}^{\infty} \psi_j^{(t)} Z_{t-j},$$

where

$$\psi_j^{(t)} = \psi_j - \frac{\sum_{k:r(k)=r(t)} \psi_{k-t+j}}{N_{r(t)}} - \frac{\sum_{l:v(l)=v(t)} \psi_{l-t+j}}{N/T} + \frac{\sum_{u=1}^N \psi_{u-t+j}}{N}. \quad (23)$$

Here, $\psi_j = 0$ when $j < 0$, implying that $\psi_j^{(t)} = 0$ if $j < t - N$.

The asymptotic limit of the sample autocovariances can now be derived:

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{N} \sum_{t=h+1}^N \epsilon_t^{\text{ols}} \epsilon_{t-h}^{\text{ols}} = \frac{1}{N} \sum_{t=h+1}^N (W_t + \delta_t)(W_{t-h} + \delta_{t-h}) \\ &= \frac{1}{N} \sum_{t=h+1}^N (W_t W_{t-h} + \delta_{t-h} W_t + \delta_t W_{t-h} + \delta_t \delta_{t-h}). \end{aligned} \quad (24)$$

Arguing as in Proposition 7.3.5 of [Brockwell and Davis \(1991, pp. 232\)](#) gives

$$\frac{1}{N} \sum_{t=h+1}^N W_t W_{t-h} = \frac{1}{N} \sum_{t=h+1}^N \sum_{j=-\infty}^{\infty} \psi_j^{(t)} \psi_{j-h}^{(t-h)} Z_{t-j}^2 + O_P \left(\frac{1}{\sqrt{N}} \right).$$

In (23), since $\sum_{j=0}^{\infty} |\psi_j| < \infty$, and $N_{r(t)} = O(N)$ for all $t = 1, \dots, N$, it is not difficult to show that there exists a positive constant c such that,

$$\sup_{t,j} \left| \psi_j^{(t)} - \psi_j \right| = \frac{c}{N}.$$

Therefore, for each t and h , $\left\{ \psi_j^{(t)} \psi_{j-h}^{(t-h)} \right\}_{j=-\infty}^{\infty}$ is absolutely convergent, and

$$\left| \sum_{j=-\infty}^{\infty} \psi_j^{(t)} \psi_{j-h}^{(t-h)} - \sum_{j=-\infty}^{\infty} \psi_j \psi_{j-h} \right| = O \left(\frac{1}{N} \right).$$

Since $\{Z_t\}$ is i.i.d., with $E[Z_t^2] = \sigma^2$, the weak law of large numbers (WLLN) for linear processes ([Brockwell and Davis 1991, pp. 208, Proposition 6.3.10](#)) gives

$$\begin{aligned} \frac{1}{N} \sum_{t=h+1}^N W_t W_{t-h} &= \frac{1}{N} \sum_{t=h+1}^N \sum_{j=-\infty}^{\infty} \psi_j^{(t)} \psi_{j-h}^{(t-h)} \sigma^2 + O_P \left(\frac{1}{\sqrt{N}} \right) \\ &= \frac{1}{N} \sum_{t=h+1}^N \sum_{j=-\infty}^{\infty} \psi_j \psi_{j-h} \sigma^2 + O_P \left(\frac{1}{\sqrt{N}} \right). \end{aligned}$$

Now using that $\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j-h}$ gives

$$\frac{1}{N} \sum_{t=h+1}^N W_t W_{t-h} = \frac{N-h}{N} \gamma(h) + O_P \left(\frac{1}{\sqrt{N}} \right) = \gamma(h) + O_P \left(\frac{1}{\sqrt{N}} \right).$$

This identifies the limit of the first term in the bottom line of (24). By (23), it is not hard to show that for each t , $\left\{ \psi_j^{(t)} \right\}_{j=-\infty}^{\infty}$ is absolutely convergent. Hence, for the second and third terms of (24), apply the WLLN again to see that these terms converge to zero in probability

at rate $O_P(1/\sqrt{N})$. Hence, as $N \rightarrow \infty$,

$$\hat{\gamma}(h) = \gamma(h) + \frac{1}{N} \sum_{t=h+1}^N \delta_t \delta_{t-h} + O_P\left(\frac{1}{\sqrt{N}}\right) = \gamma(h) + \delta^2 + O_P\left(\frac{1}{\sqrt{N}}\right),$$

which proves (22). \square

A.2 Proof of asymptotic consistency of the univariate BMDL

To simplify the BMDL formula (14), we first establish an asymptotic result for $\hat{\sigma}^2$. Recall that BMDL and MDL estimators are respectively

$$\hat{\sigma}^2 = \frac{1}{N-p} \hat{\mathbf{X}}' \left[\hat{\mathbf{B}} - \hat{\mathbf{B}} \hat{\mathbf{A}} \left(\hat{\mathbf{A}}' \hat{\mathbf{B}} \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}' \hat{\mathbf{B}} \right] \hat{\mathbf{X}}, \quad \hat{\sigma}_{\nu=\infty}^2 = \frac{1}{N-p} \hat{\mathbf{X}}' \left(\mathbf{I}_N - \mathcal{P}_{[\hat{\mathbf{A}} \ \hat{\mathbf{D}}]} \right) \hat{\mathbf{X}}.$$

Lemma 2. *Under any changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, as $N \rightarrow \infty$,*

$$\hat{\sigma}^2 = \hat{\sigma}_{\nu=\infty}^2 + O_P\left(\frac{1}{N}\right). \quad (25)$$

Proof. Since $\hat{\boldsymbol{\phi}}$ has the limit in (22), it is not hard to show that as N tends to infinity, $\hat{\mathbf{D}}' \hat{\mathbf{D}}/N$ and $\hat{\mathbf{D}}' \hat{\mathbf{X}}/N$ converges in probability to a $m \times m$ positive definite matrix and an m -dimensional vector, respectively, both at rates $O_P(1/N)$. In the prior of $\boldsymbol{\mu}$, the parameter ν is a constant; hence,

$$\begin{aligned} \frac{1}{N} \hat{\mathbf{X}}' \hat{\mathbf{B}} \hat{\mathbf{X}} &= \frac{\hat{\mathbf{X}}' \hat{\mathbf{X}}}{N} - \frac{\hat{\mathbf{X}}' \hat{\mathbf{D}}}{N} \left(\frac{\hat{\mathbf{D}}' \hat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu} \right)^{-1} \frac{\hat{\mathbf{D}}' \hat{\mathbf{X}}}{N} \\ &= \frac{\hat{\mathbf{X}}' \hat{\mathbf{X}}}{N} - \frac{\hat{\mathbf{X}}' \hat{\mathbf{D}}}{N} \left(\frac{\hat{\mathbf{D}}' \hat{\mathbf{D}}}{N} \right)^{-1} \frac{\hat{\mathbf{D}}' \hat{\mathbf{X}}}{N} + O_P\left(\frac{1}{N}\right) = \frac{1}{N} \hat{\mathbf{X}}' (\mathbf{I}_{N-p} - \mathcal{P}_{\hat{\mathbf{D}}}) \hat{\mathbf{X}} + O_P\left(\frac{1}{N}\right). \end{aligned}$$

Similar arguments give

$$\frac{1}{N} \hat{\mathbf{X}}' \hat{\mathbf{B}} \hat{\mathbf{A}} = \frac{1}{N} \hat{\mathbf{X}}' (\mathbf{I}_{N-p} - \mathcal{P}_{\hat{\mathbf{D}}}) \hat{\mathbf{A}} + O_P\left(\frac{1}{N}\right), \quad \frac{1}{N} \hat{\mathbf{A}}' \hat{\mathbf{B}} \hat{\mathbf{A}} = \frac{1}{N} \hat{\mathbf{A}}' (\mathbf{I}_{N-p} - \mathcal{P}_{\hat{\mathbf{D}}}) \hat{\mathbf{A}} + O_P\left(\frac{1}{N}\right).$$

Hence, the left hand side of (25) has the limit

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N-p} \hat{\mathbf{X}}' \left[\hat{\mathbf{B}} - \hat{\mathbf{B}} \hat{\mathbf{A}} \left(\hat{\mathbf{A}}' \hat{\mathbf{B}} \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}' \hat{\mathbf{B}} \right] \hat{\mathbf{X}} \\
&= \frac{1}{N-p} \hat{\mathbf{X}}' \left[\mathbf{I} - \mathcal{P}_{\hat{\mathbf{D}}} - \mathcal{P}_{(\mathbf{I}_{N-p} - \mathcal{P}_{\hat{\mathbf{D}}}) \hat{\mathbf{A}}} \right] \hat{\mathbf{X}} + O_P \left(\frac{1}{N} \right) \\
&= \frac{1}{N-p} \hat{\mathbf{X}}' \left(\mathbf{I}_{N-p} - \mathcal{P}_{[\hat{\mathbf{A}} \ \hat{\mathbf{D}}]} \right) \hat{\mathbf{X}} + O_P \left(\frac{1}{N} \right) = \hat{\sigma}_{\nu=\infty}^2 + O_P \left(\frac{1}{N} \right),
\end{aligned}$$

where the second to last equality follows by (17). \square

Lemma 3. *Define a function*

$$f(\delta^2) = \gamma(0) + \delta^2 - (\gamma_p + \delta^2 \mathbf{1}_p)' (\boldsymbol{\Gamma}_p + \delta^2 \mathbf{J}_p)^{-1} (\gamma_p + \delta^2 \mathbf{1}_p). \quad (26)$$

Then $f(\delta^2)$ is strictly increasing in δ^2 , and under any changepoint model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$,

$$\hat{\sigma}^2 = f(\delta^2) + O_P \left(\frac{1}{\sqrt{N}} \right). \quad (27)$$

Proof. We first show that for any $\boldsymbol{\lambda}$ with $m > 0$,

$$\frac{1}{N} \hat{\mathbf{X}}' \left(\mathbf{I}_N - \mathcal{P}_{[\hat{\mathbf{A}} \ \hat{\mathbf{D}}]} \right) \hat{\mathbf{X}} = \hat{\gamma}(0) - \hat{\gamma}_p' \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\gamma}_p + O_P \left(\frac{1}{N} \right). \quad (28)$$

For notational simplicity, for any $j = 0, 1, \dots, p$, matrices formed by the rows of \mathbf{A} and \mathbf{D} are denoted by

$$\mathbf{A}_j \stackrel{\text{def}}{=} \mathbf{A}_{(p+1-j):(N-j)}, \quad \mathbf{D}_j \stackrel{\text{def}}{=} \mathbf{D}_{(p+1-j):(N-j)}.$$

Since both $\hat{\mathbf{A}}$ and \mathbf{A}_j are $(N-p) \times T$ matrices and each column in $\hat{\mathbf{A}}$ can be written as a linear combination of the columns in \mathbf{A}_j , the corresponding column spaces agree: $C(\hat{\mathbf{A}}) = C(\mathbf{A}_j)$. Therefore, $\mathcal{P}_{\hat{\mathbf{A}}} = \mathcal{P}_{\mathbf{A}_j}$ for all j . Now define

$$\boldsymbol{\Delta}_j = \mathbf{D}_j - \frac{\hat{\mathbf{D}}}{1 - \hat{\phi}_1 - \hat{\phi}_2 - \dots - \hat{\phi}_p}. \quad (29)$$

The denominator in (29) cannot be zero since $1 - \sum_{k=1}^p \hat{\phi}_k \neq 0$ for any Yule-Walker estimates (Brockwell and Davis 1991).

Since there are at most $2m(p+h)$ non-zero entries in $\mathbf{\Delta}_j$, and none of these entries depend on N , $\mathbf{\Delta}_j' \mathbf{\Delta}_j = O_P(1)$. In addition, for any N -dimensional vectors $\boldsymbol{\alpha}$ whose entries do not depend on N , $\boldsymbol{\alpha}' \mathbf{\Delta}_j = O_P(1)$. Using (29),

$$\begin{aligned} \frac{\hat{\mathbf{D}}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \hat{\mathbf{D}}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k\right)^2} &= \frac{1}{N} (\mathbf{D}_j - \mathbf{\Delta}_j)' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) (\mathbf{D}_j - \mathbf{\Delta}_j) = \frac{\mathbf{D}_j' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \mathbf{D}_j}{N} + O_P\left(\frac{1}{N}\right), \\ \frac{\boldsymbol{\alpha}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \hat{\mathbf{D}}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k\right)} &= \frac{1}{N} \boldsymbol{\alpha}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) (\mathbf{D}_j - \mathbf{\Delta}_j) = \frac{\boldsymbol{\alpha}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \mathbf{D}_j}{N} + O_P\left(\frac{1}{N}\right). \end{aligned}$$

Therefore, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^N$ whose entries do not depend on N ,

$$\begin{aligned} \frac{1}{N} \boldsymbol{\alpha}' \mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \hat{\mathbf{D}}} \boldsymbol{\beta} &= \frac{\boldsymbol{\alpha}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \hat{\mathbf{D}}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k\right)} \left(\frac{\hat{\mathbf{D}}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \hat{\mathbf{D}}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k\right)^2} \right)^{-1} \frac{\hat{\mathbf{D}}' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \boldsymbol{\beta}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k\right)} \\ &= \frac{1}{N} \boldsymbol{\alpha}' \left[(\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \mathbf{D}_j (\mathbf{D}_j' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \mathbf{D}_j)^{-1} \mathbf{D}_j' (\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \right] \boldsymbol{\beta} + O_P\left(\frac{1}{N}\right) \\ &= \frac{1}{N} \boldsymbol{\alpha}' \mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\hat{\mathbf{A}}}) \mathbf{D}_j} \boldsymbol{\beta} + O_P\left(\frac{1}{N}\right). \end{aligned}$$

Hence, from (17),

$$\frac{1}{N} \boldsymbol{\alpha}' \mathcal{P}_{[\hat{\mathbf{A}} \ \hat{\mathbf{D}}]} \boldsymbol{\beta} = \frac{1}{N} \boldsymbol{\alpha}' \mathcal{P}_{[\mathbf{A}_j \ \mathbf{D}_j]} \boldsymbol{\beta} + O_P\left(\frac{1}{N}\right). \quad (30)$$

Since $\hat{\mathbf{X}} = \mathbf{X}_{(p+1):N} - \sum_{j=1}^p \hat{\phi}_j \mathbf{X}_{(p+1-j):(N-j)}$, for any $j, k \in \{0, 1, \dots, p\}$, (30) shows that

$$\begin{aligned} &\frac{1}{N} \mathbf{X}_{(p+1-j):(N-j)}' \left(\mathbf{I}_N - \mathcal{P}_{[\hat{\mathbf{A}} \ \hat{\mathbf{D}}]} \right) \mathbf{X}_{(p+1-k):(N-k)} \\ &= \frac{1}{N} \left[(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_j \ \mathbf{D}_j]}) \mathbf{X}_{(p+1-j):(N-j)} \right]' \left[(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_k \ \mathbf{D}_k]}) \mathbf{X}_{(p+1-k):(N-k)} \right] + O_P\left(\frac{1}{N}\right) \\ &= \frac{1}{N} \left(\boldsymbol{\epsilon}_{(p+1-j):(N-j)}^{\text{ols}} \right)' \boldsymbol{\epsilon}_{(p+1-k):(N-k)}^{\text{ols}} + O_P\left(\frac{1}{N}\right). \end{aligned}$$

Therefore, the left hand side of (28) is

$$\begin{aligned}
& \frac{1}{N} \widehat{\mathbf{X}}' \left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}} \ \widehat{\mathbf{D}}]} \right) \widehat{\mathbf{X}} \\
&= \frac{1}{N} \left[\boldsymbol{\epsilon}_{(p+1):N}^{\text{ols}} - \sum_{j=1}^p \hat{\phi}_j \boldsymbol{\epsilon}_{(p+1-j):(N-j)}^{\text{ols}} \right]' \left[\boldsymbol{\epsilon}_{(p+1):N}^{\text{ols}} - \sum_{k=1}^p \hat{\phi}_k \boldsymbol{\epsilon}_{(p+1-k):(N-k)}^{\text{ols}} \right] + O_P \left(\frac{1}{N} \right) \\
&= \hat{\gamma}(0) - 2 \sum_{j=1}^p \hat{\phi}_j \hat{\gamma}(j) + \sum_{j=1}^p \sum_{k=1}^p \hat{\phi}_j \hat{\phi}_k \hat{\gamma}(|j-k|) + O_P \left(\frac{1}{N} \right) \\
&= \hat{\gamma}(0) - 2 \hat{\gamma}'_p \hat{\phi} + \hat{\phi}' \hat{\Gamma}_p \hat{\phi} + O_P \left(\frac{1}{N} \right) \\
&= \hat{\gamma}(0) - \hat{\gamma}'_p \hat{\Gamma}_p^{-1} \hat{\gamma}_p + O_P \left(\frac{1}{N} \right),
\end{aligned}$$

which is the right hand side of (28). It is not hard to show that under the model $\boldsymbol{\lambda}_\emptyset$ ($m=0$), because $C(\widehat{\mathbf{D}})$ is the null space, (28) also holds.

Note that the left hand side of (28) is $(N-p)\hat{\sigma}_{\nu=\infty}^2/N$, so under large N , it is $\hat{\sigma}_{\nu=\infty}^2 + O_P(1/N)$, and by Lemma 2, equivalently to $\hat{\sigma}^2 + O_P(1/N)$. By Proposition 2, the right hand side of (28) is $f(\delta^2) + O_P(1/\sqrt{N})$; hence, (27) holds.

We next show that $f(\delta^2)$ in (26) is strictly increasing in δ^2 . If $\delta^2 > 0$, according to (2.22) in Harville (2008, pp. 428), for any matrices $\mathbf{R} \in \mathbb{R}^{r \times r}$, $\mathbf{S} \in \mathbb{R}^{r \times l}$, $\mathbf{T} \in \mathbb{R}^{l \times l}$, $\mathbf{U} \in \mathbb{R}^{l \times r}$ with \mathbf{R}, \mathbf{U} non-singular, $(\mathbf{R} + \mathbf{S}\mathbf{T}\mathbf{U})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{S}(\mathbf{T}^{-1} + \mathbf{U}\mathbf{R}^{-1}\mathbf{S})^{-1}\mathbf{U}\mathbf{R}^{-1}$. Hence,

$$(\Gamma_p + \delta^2 \mathbf{J}_p)^{-1} = (\Gamma_p + \mathbf{1}_p \delta^2 \mathbf{1}_p')^{-1} = \Gamma_p^{-1} - \Gamma_p^{-1} \mathbf{1}_p \left(\frac{1}{\delta^2} + \mathbf{1}_p' \Gamma_p^{-1} \mathbf{1}_p \right)^{-1} \mathbf{1}_p' \Gamma_p^{-1}. \quad (31)$$

For notational simplicity, denote the following scalars by

$$a \stackrel{\text{def}}{=} \mathbf{1}_p' \Gamma_p^{-1} \mathbf{1}_p, \quad b \stackrel{\text{def}}{=} \mathbf{1}_p' \Gamma_p^{-1} \gamma_p. \quad (32)$$

Then $f(\delta^2)$ can be expanded as

$$f(\delta^2) = \gamma(0) + \delta^2 - \gamma_p' \Gamma_p^{-1} \gamma_p - 2b\delta^2 - a(\delta^2)^2 + \frac{b^2}{\frac{1}{\delta^2} + a} + \frac{2ab\delta^2}{\frac{1}{\delta^2} + a} + \frac{a^2(\delta^2)^2}{\frac{1}{\delta^2} + a}.$$

Differentiation of $f(\delta^2)$ with respect to δ^2 gives

$$f'(\delta^2) = 1 - 2b - 2a\delta^2 + \frac{b^2 \frac{1}{(\delta^2)^2}}{\left(\frac{1}{\delta^2} + a\right)^2} + \frac{2ab \left(\frac{2}{\delta^2} + a\right)}{\left(\frac{1}{\delta^2} + a\right)^2} + \frac{a^2 (3 + 2a\delta^2)}{\left(\frac{1}{\delta^2} + a\right)^2} = \frac{(b-1)^2}{(1+a\delta^2)^2} > 0.$$

The last inequality follows since $\{\epsilon_t\}_{t=1}^N$ is causal, which implies that $b = \sum_{k=1}^p \phi_k > 1$. Therefore, $f(\delta^2)$ is strictly increasing in δ^2 . □

Lemma 4. *Under any changepoint model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ with $m > 0$,*

$$\frac{1}{2} \log \left(\left| \widehat{\mathbf{D}}' \widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right| \right) = \frac{1}{2} \sum_{j=2}^{m+1} \log(N_r) - m \log \left(1 - \sum_{k=1}^p \hat{\phi}_k \right) + O_P \left(\frac{1}{N} \right). \quad (33)$$

Proof. By (29) and the corresponding results in the proof of Lemma 3, as $N \rightarrow \infty$,

$$\frac{\widehat{\mathbf{D}}' \widehat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu} = \frac{\widehat{\mathbf{D}}' \widehat{\mathbf{D}}}{N} + O \left(\frac{1}{N} \right) = \frac{\mathbf{D}' \mathbf{D}}{N \left(1 - \sum_{k=1}^p \hat{\phi}_k \right)^2} + O_P \left(\frac{1}{N} \right).$$

The determinant of the $m \times m$ matrix (of finite dimension) is then

$$\begin{aligned} \log \left(\left| \widehat{\mathbf{D}}' \widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right| \right) &= m \log(N) + \log \left(\left| \frac{\widehat{\mathbf{D}}' \widehat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu} \right| \right) \\ &= m \log(N) + \log \left(\frac{|\mathbf{D}' \mathbf{D}|}{N^m \left(1 - \sum_{k=1}^p \hat{\phi}_k \right)^{2m}} \right) + O_P \left(\frac{1}{N} \right) \\ &= \log(|\mathbf{D}' \mathbf{D}|) - 2m \log \left(1 - \sum_{k=1}^p \hat{\phi}_k \right) + O_P \left(\frac{1}{N} \right) \\ &= \log \left(\prod_{j=2}^{m+1} N_r \right) - 2m \log \left(1 - \sum_{k=1}^p \hat{\phi}_k \right) + O_P \left(\frac{1}{N} \right), \end{aligned}$$

and (33) follows immediately. □

The asymptotic consistency of the BMDL can now be proven.

Proof of Theorem 1. For any changepoint model $\boldsymbol{\lambda}$ with $m > 0$, since under infill asymptotics, $N_r = O(N)$ for all $r = 2, \dots, m+1$, by Lemma 4,

$$\frac{1}{2} \log \left(\left| \widehat{\mathbf{D}}' \widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right| \right) = \frac{m}{2} \log(N) + O_P(1).$$

Therefore, by (27), the asymptotic BMDL (14) for the model $\boldsymbol{\lambda}$ (including $\boldsymbol{\lambda}_\phi$) is

$$\begin{aligned} \text{BMDL}(\boldsymbol{\lambda}) &= \frac{N-p}{2} \log \left[f(\delta^2) + O_P \left(\frac{1}{\sqrt{N}} \right) \right] + \frac{m}{2} \log(N) \\ &\quad - \sum_{k=1}^2 \log \left[\Gamma(a + m^{(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{(k)}) \right] + O_P(1). \end{aligned}$$

It now follows that the difference between BMDLs in a (non-true) model $\boldsymbol{\lambda}$ and the true model $\boldsymbol{\lambda}^0$ is asymptotically

$$\begin{aligned} \text{BMDL}(\boldsymbol{\lambda}) - \text{BMDL}(\boldsymbol{\lambda}^0) &= \frac{N-p}{2} \log \left[\frac{f(\delta^2) + O_P \left(\frac{1}{\sqrt{N}} \right)}{f(0) + O_P \left(\frac{1}{\sqrt{N}} \right)} \right] + \frac{m - m^0}{2} \log(N) \\ &\quad + \sum_{k=1}^2 \log \left[\frac{\Gamma(a + m^{0(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{0(k)})}{\Gamma(a + m^{(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{(k)})} \right] + O_P(1). \end{aligned} \quad (34)$$

Since $\delta^2 = 0$ if and only if $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$, as $N \rightarrow \infty$, the first term of (34) is

$$\frac{N-p}{2} \log \left[\frac{f(\delta^2) + O_P \left(\frac{1}{\sqrt{N}} \right)}{f(0) + O_P \left(\frac{1}{\sqrt{N}} \right)} \right] = \begin{cases} O_P(N) > 0, & \text{if } \boldsymbol{\lambda} \not\supset \boldsymbol{\lambda}^0, \\ O_P(1), & \text{if } \boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0. \end{cases}$$

Without loss of generality, the number of documented and undocumented times are assumed to be increasing linearly with N , say $N^{(k)} = O(N)$, for $k = 1, 2$. Stirling's formula

allows us to find the asymptotic limit of Gamma function ratios:

$$\begin{aligned} \frac{\Gamma(b^{(k)} + N^{(k)} - m^{0(k)})}{\Gamma(b^{(k)} + N^{(k)} - m^{(k)})} &\approx e^{m^{0(k)} - m^{(k)}} \frac{(b^{(k)} + N^{(k)} - m^{0(k)} - 1)^{b^{(k)} + N^{(k)} - m^{0(k)} - 1/2}}{(b^{(k)} + N^{(k)} - m^{(k)} - 1)^{b^{(k)} + N^{(k)} - m^{(k)} - 1/2}} \\ &= O\left(N^{m^{(k)} - m^{0(k)}}\right). \end{aligned}$$

Therefore, in (34),

$$\sum_{k=1}^2 \log \left[\frac{\Gamma(a + m^{0(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{0(k)})}{\Gamma(a + m^{(k)}) \Gamma(b^{(k)} + N^{(k)} - m^{(k)})} \right] = (m - m^0) \log N + O_P(1).$$

If $\lambda \not\supset \lambda^0$, the first term in (34) is asymptotically dominant:

$$\text{BMDL}(\lambda) - \text{BMDL}(\lambda^0) = O_P(N) + 1.5(m - m^0) \log N = O_P(N) > 0.$$

In contrast, if $\lambda \supset \lambda^0$, then since $m > m^0$,

$$\text{BMDL}(\lambda) - \text{BMDL}(\lambda^0) = O_P(1) + 1.5(m - m^0) \log N = O_P(\log N) > 0.$$

□

References

- Aue, A., Cheung, R. C. Y., Lee, T. C. M., and Zhong, M. (2014), “Segmented Model Selection in Quantile Regression Using the Minimum Description Length Principle,” *Journal of the American Statistical Association*, 109, 1241–1256.
- Barry, D. and Hartigan, J. A. (1993), “A Bayesian Analysis for Change Point Problems,” *Journal of the American Statistical Association*, 88, 309–319.
- Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, Springer-Verlag, 2nd ed.
- Carlin, B. P. and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall/CRC Boca Raton.
- Caussinus, H. and Mestre, O. (2004), “Detection and Correction of Artificial Shifts in Climate Series,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53, 405–425.
- Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014), “Group LASSO for Structural Break Time Series,” *Journal of the American Statistical Association*, 109, 590–599.

- Chernoff, H. and Zacks, S. (1964), “Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time,” *The Annals of Mathematical Statistics*, 35, 999–1018.
- Chib, S. (1998), “Estimation and Comparison of Multiple Change-point Models,” *Journal of Econometrics*, 86, 221–241.
- Christensen, R. (2002), *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer.
- Clyde, M. A. and George, E. I. (2004), “Model Uncertainty,” *Statistical Science*, 19, 81–94.
- Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), “Structural Break Estimation for Nonstationary Time Series Models,” *Journal of the American Statistical Association*, 101, 223–239.
- (2008), “Break Detection for a Class of Nonlinear Time Series Models,” *Journal of Time Series Analysis*, 29, 834–867.
- Davis, R. A. and Yau, C. Y. (2013), “Consistency of Minimum Description Length Model Selection for Piecewise Stationary Time Series Models,” *Electronic Journal of Statistics*, 7, 381–411.
- Du, C., Kao, C.-L. M., and Kou, S. C. (2015), “Stepwise Signal Extraction via Marginal Likelihood,” *Journal of the American Statistical Association*, in press.
- Fearnhead, P. and Vasileiou, D. (2009), “Bayesian Analysis of Isochores,” *Journal of the American Statistical Association*, 104, 132–141.
- García-Donato, G. and Martínez-Beneito, M. A. (2013), “On Sampling Strategies in Bayesian Variable Selection Problems with Large Model Spaces,” *Journal of the American Statistical Association*, 108, 340–352.
- George, E. I. and McCulloch, R. E. (1997), “Approaches for Bayesian Variable Selection,” *Statistics Sinica*, 7, 339–373.
- Giordani, P. and Kohn, R. (2008), “Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models,” *Journal of Business and Economic Statistics*, 26, 66–77.
- Girón, J., Moreno, E., and Casella, G. (2007), “Objective Bayesian Analysis of Multiple Change-points for Linear Models,” *Bayesian Statistics 8*.
- Green, Peter, J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732.
- Hannart, A. and Naveau, P. (2012), “An Improved Bayesian Information Criterion for Multiple Change-point Models,” *Technometrics*, 54, 256–268.
- Hansen, M. H. and Yu, B. (2001), “Model Selection and the Principle of Minimum Description Length,” *Journal of the American Statistical Association*, 96, 746–774.
- Harville, D. A. (2008), *Matrix Algebra From a Statistician’s Perspective*, Springer-Verlag.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes Factors,” *Journal of the American Statistical Association*, 90, 773–795.
- Killick, R., Fearnhead, P., and Eckley, I. A. (2012), “Optimal Detection of Change-points With a Linear Computational Cost,” *Journal of the American Statistical Association*, 107, 1590–1598.

- Lee, T. C. M. (2000), “A Minimum Description Length-Based Image Segmentation Procedure, and its Comparison with a Cross-Validation-Based Segmentation Procedure,” *Journal of the American Statistical Association*, 95, 259–270.
- Li, F. and Zhang, N. R. (2010), “Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics,” *Journal of the American Statistical Association*, 105, 1202–1214.
- Li, S. and Lund, R. (2012), “Multiple Changepoint Detection via Genetic Algorithms,” *Journal of Climate*, 25, 674–686.
- Li, Y. and Lund, R. (2015), “Multiple Changepoint Detection Using Metadata,” *Journal of Climate*, 28, 4199–4216.
- Liu, G., Shao, Q., Lund, R., and Woody, J. (2015), “Testing for Seasonal Means in Time Series Data,” *In press, Environmetrics*.
- Lu, Q., Lund, R., and Lee, T. C. M. (2010), “An MDL Approach to the Climate Segmentation Problem,” *The Annals of Applied Statistics*, 4, 299–319.
- Lu, Q. Q. and Lund, R. (2007), “Simple Linear Regression with Multiple Level Shifts,” *Canadian Journal of Statistics*, 37, 447–458.
- Menne, M. J. and Williams Jr, C. N. (2005), “Detection of Undocumented Changepoints Using Multiple Test Statistics and Composite Reference Series,” *Journal of Climate*, 18, 4271–4286.
- (2009), “Homogenization of Temperature Series via Pairwise Comparisons,” *Journal of Climate*, 22, 1700–1717.
- Mitchell, J. M. (1953), “On the Causes of Instrumentally Observed Secular Temperature Trends,” *Journal of Meteorology*, 10, 244–261.
- Preuss, P., Puchstein, R., and Dette, H. (2015), “Detection of Multiple Structural Breaks in Multivariate Time Series,” *Journal of the American Statistical Association*, 110, 654–668.
- Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, vol. 511, World Scientific, Singapore.
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.
- Scott, J. and Berger, J. (2010), “Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-selection Problem,” *The Annals of Statistics*, 38, 2587–2619.
- Shannon, C. E. (1948), “A Mathematical Theory of Communication,” *Bell System Technical Journal*, 27, 623.
- Shao, X. and Zhang, X. (2010), “Testing for Change Points in Time Series,” *Journal of the American Statistical Association*, 105, 1228–1240.
- Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences*, Academic Press.
- Yao, Y.-C. (1984), “Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches,” *The Annals of Statistics*, 12, 1434–1447.
- Yau, C. Y., Tang, C. M., and Lee, T. C. M. (2015), “Estimation of Multiple-Regime Threshold Autoregressive Models with Structural Breaks,” *Journal of the American Statistical Association*, 110, 1175–1186.
- Yau, C. Y. and Zhao, Z. (2015), “Inference for Multiple Change Points in Time Series via Likelihood Ratio Scan Statistics,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.